Unit 9: Introduction to Multiple Regression and Statistical Interaction

Unit 9 Post Hole:

Interpret the parameter estimates and F-test from regressing a continuous variable on a set of dummy variables.

Unit 9 Technical Memo and School Board Memo:

Regress a continuous variable on a polychotomous variable, fit the equivalent one-way ANOVA model, produce appropriate tables and discuss your results.

Unit 9 (and Unit 10) Reading:

http://onlinestatbook.com/
Chapter 13, ANOVA

Unit 9: Technical Memo and School Board Memo

Work Products (Part I of II):

I. Technical Memo: Have one section per biviariate analysis. For each section, follow this outline. (4 Sections)

   A. Introduction

      i. State a theory (or perhaps hunch) for the relationship—think causally, be creative. (1 Sentence)

      ii. State a research question for each theory (or hunch)—think correlationally, be formal. Now that you know the statistical machinery that justifies an inference from a sample to a population, begin each research question, “In the population,…” (1 Sentence)

      iii. List the two variables, and label them “outcome” and “predictor,” respectively.

      iv. Include your theoretical model.

   B. Univariate Statistics. Describe your variables, using descriptive statistics. What do they represent or measure?

      i. Describe the data set. (1 Sentence)

      ii. Describe your variables. (1 Short Paragraph Each)

         a. Define the variable (parenthetically noting the mean and s.d. as descriptive statistics).

         b. Interpret the mean and standard deviation in such a way that your audience begins to form a picture of the way the world is. Never lose sight of the substantive meaning of the numbers.

         c. Polish off the interpretation by discussing whether the mean and standard deviation can be misleading, referencing the median, outliers and/or skew as appropriate.

   C. Correlations. Provide an overview of the relationships between your variables using descriptive statistics.

      i. Interpret all the correlations with your outcome variable. Compare and contrast the correlations in order to ground your analysis in substance. (1 Paragraph)

      ii. Interpret the correlations among your predictors. Discuss the implications for your theory. As much as possible, tell a coherent story. (1 Paragraph)

      iii. As you narrate, note any concerns regarding assumptions (e.g., outliers or non-linearity), and, if a correlation is uninterpretable because of an assumption violation, then do not interpret it.

Unit 9: Technical Memo and School Board Memo

Work Products (Part II of II):

I. Technical Memo (continued)

   D. Regression Analysis. Answer your research question using inferential statistics. (1 Paragraph)

      i. Include your fitted model.

      ii. Use the R^2 statistic to convey the goodness of fit for the model (i.e., strength).

      iii. To determine statistical significance, test the null hypothesis that the magnitude in the population is zero, reject (or not) the null hypothesis, and draw a conclusion (or not) from the sample to the population.

      iv. Describe the direction and magnitude of the relationship in your sample, preferably with illustrative examples. Draw out the substance of your findings through your narrative.

      v. Use confidence intervals to describe the precision of your magnitude estimates so that you can discuss the magnitude in the population.

      vi. If simple linear regression is inappropriate, then say so, briefly explain why, and forego any misleading analysis.

   X. Exploratory Data Analysis. Explore your data using outlier resistant statistics.

      i. For each variable, use a coherent narrative to convey the results of your exploratory univariate analysis of the data. Don’t lose sight of the substantive meaning of the numbers. (1 Paragraph Each)

      ii. For the relationship between your outcome and predictor, use a coherent narrative to convey the results of your exploratory bivariate analysis of the data. (1 Paragraph)

II. School Board Memo: Concisely, precisely and plainly convey your key findings to a lay audience. Note that, whereas you are building on the technical memo for most of the semester, your school board memo is fresh each week. (Max 200 Words)

III. Memo Metacognitive

Unit 9: Road Map (VERBAL)

Nationally Representative Sample of 7,800 8th Graders Surveyed in 1988 (NELS 88).
Outcome Variable (aka Dependent Variable):
READING, a continuous variable, test score, mean = 47 and standard deviation = 9
Predictor Variables (aka Independent Variables):
FREELUNCH, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
RACE, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White

- Unit 1: In our sample, is there a relationship between reading achievement and free lunch?
- Unit 2: In our sample, what does reading achievement look like (from an outlier resistant perspective)?
- Unit 3: In our sample, what does reading achievement look like (from an outlier sensitive perspective)?
- Unit 4: In our sample, how strong is the relationship between reading achievement and free lunch?
- Unit 5: In our sample, free lunch predicts what proportion of variation in reading achievement?
- Unit 6: In the population, is there a relationship between reading achievement and free lunch?
- Unit 7: In the population, what is the magnitude of the relationship between reading and free lunch?
- Unit 8: What assumptions underlie our inference from the sample to the population?
- Unit 9: In the population, is there a relationship between reading and race?
- Unit 10: In the population, is there a relationship between reading and race controlling for free lunch?
- Appendix A: In the population, is there a relationship between race and free lunch?

Unit 9: Roadmap (R Output)

```
> load("E:/User/Folder/RoadmapData.rda")
> library(abind, pos=4)
> numSummary(RoadmapData[,c("FREELUNCH", "READING")],
+   statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
               mean       sd    0%   25%   50%   75%  100%    n
FREELUNCH 0.3353846 0.472155  0.00  0.00  0.00  1.00  1.00 7800
READING  47.4940397 8.569440 23.96 41.24 47.43 53.93 63.49 7800
```

Unit 3 (mean, sd) · Unit 2 (quantiles, n)

```
> RegModel.1 <- lm(READING~FREELUNCH, data=RoadmapData)
> summary(RegModel.1, cor=FALSE)
Call:
lm(formula = READING ~ FREELUNCH, data = RoadmapData)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.1176     0.1147  428.17   <2e-16 ***
FREELUNCH    -4.8409     0.1981  -24.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.26 on 7798 degrees of freedom
Multiple R-squared: 0.07114, Adjusted R-squared: 0.07102
F-statistic: 597.3 on 1 and 7798 DF,  p-value: < 2.2e-16
```

Unit 1 (Estimate) · Unit 8 (Std. Error) · Unit 6 (t value, Pr(>|t|)) · Unit 5 (Adjusted R-squared) · Unit 9 (F-statistic)

```
> library(MASS, pos=4)
> Confint(RegModel.1, level=.95)
             Estimate     2.5 %    97.5 %
(Intercept) 49.117616 48.892742 49.342489
FREELUNCH   -4.840938 -5.229237 -4.452638
```

Unit 7

```
> cor(RoadmapData[,c("FREELUNCH","READING")])
           FREELUNCH    READING
FREELUNCH  1.0000000 -0.2667237
READING   -0.2667237  1.0000000
```

Unit 4

Unit 9: Roadmap (SPSS Output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.267[a]	.071	.071	8.25952

a. Predictors: (Constant), FREELUNCH

Unit 5

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	40744.322	1	40744.322	597.251	.000[a]
	Residual	531977.541	7798	68.220		
	Total	572721.864	7799			

a. Predictors: (Constant), FREELUNCH

b. Dependent Variable: READING

Unit 9

Statistics

		READING	FREELUNCH
N	Valid	7800	7800
	Missing	0	0
Mean		47.4940	.3354
Std. Deviation		8.56944	.47216
Minimum		23.96	.00
Maximum		63.49	1.00
Percentiles	25	41.2400	.0000
	50	47.4300	.0000
	75	53.9300	1.0000

Unit 3

Unit 2

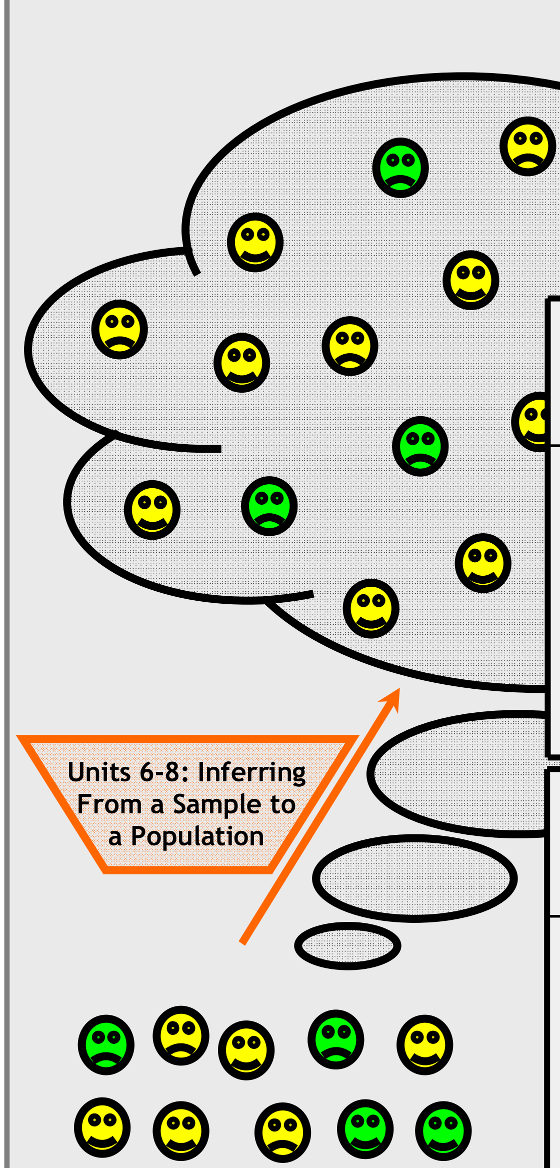
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	49.118	.115		428.169	.000	48.893	49.342
	FREELUNCH	-4.841	.198	-.267	-24.439	.000	-5.229	-4.453

a. Dependent Variable: READING

Unit 1 Unit 8 Unit 4 Unit 6 Unit 7

Unit 9: Road Map (Schematic)



		Single Predictor		
		Continuous	Polychotomous	Dichotomous
Outcome	Continuous	Regression	Regression ANOVA	Regression ANOVA T-tests
	Polychotomous	Logistic Regression	Chi Squares	Chi Squares
	Dichotomous		Chi Squares	Chi Squares

		Multiple Predictors		
		Continuous	Polychotomous	Dichotomous
Outcome	Continuous	Multiple Regression	Regression ANOVA	Regression ANOVA
	Polychotomous	Logistic Regression	Chi Squares	Chi Squares
	Dichotomous		Chi Squares	Chi Squares

Epistemological Minute

My first gig as a professional data analyst was working with a high school data team of teachers and administrators who built a school data set but did not know how to use it. I had the skills to use the data for addressing their theories and for answering their research questions, but I first needed to understand their theories and research questions. I spent the first few meetings listening and learning. When I was finally ready to dive into the data, I found that the race/ethnicity variable had 57 possible categories, 13 of which were represented in the school, and 7 of which were represented by only one or two students each. For example, only one student self-identified as "Latino & Black & Pacific Islander." For data analytic purposes, I needed to reduce the 13 categories to a few manageable categories. The technicalities were no problem, so I immediately started re-categorizing and re-binning: Asian, Black, Latino, White, Mixed… Then, when I realized what I was doing, I got sick to my stomach. Who was I to be assigning race/ethnicity classifications? I was not the right person, but who was?

I concluded that the data team should decide on the classifications; first, because the team was sensitive to the complex issues of race/ethnicity and education, and second, because the data team was going to use the results, and the results would be useless if they did not understand the results. The data team members needed to understand the classificatory system (and its arbitrariness and limitations) if they were going to use the results, and the best way for them to understand the classificatory system was for them to devise it.

For me, this was largely a question of meaning. From the philosophy of language, the source of meaning has three components: syntax, semantics and pragmatics. Each component suggests data analytic rules and responsibilities.

Syntax—In order to effectively convey meaning, language must be structured. In verbal language, we need nouns and verbs indicated as such by their forms. In mathematical language, we need equalities, operators, numbers and variables indicated as such by their forms. The edict for data analysts: Use grammar comprehensible to your audience. For example, if your audience does not understand mathematical grammar, the data analyst is responsible for either teaching the mathematics or finding a verbal (or visual) alternative to the mathematics.

Semantics-In order to effectively convey meaning, language must be meaningful. The edict for data analysts: Use only terms that your audience understands. For example, if your audience does not understand what you mean by "race/ethnicity," then explain what you mean. Define your variables with care.

Pragmatics deserves two slides of its own.

Epistemological Minute

Pragmatics—In order to effectively convey meaning, language must be adaptive to the purposes at hand.

In his *Studies in the Way of Words* (1989), Paul Grice argues that, if we are using language for the purposes of cooperation, then there are four "maxims" that we must follow:

- **The Maxim of Quantity: "Make your contribution as informative as is required (for the current purposes of the exchange). Do not make your contribution more informative than is required."**
- **The Maxim of Quality: "Try to make your contribution one that is true."**
  - **"Do not say what you believe to be false."**
  - **"Do not say that for which you lack adequate evidence."**
- **The Maxim of Relation: "Be relevant."**
- **The Maxim of Manner: "Be perspicuous."**
  - **"Avoid obscurity of expression."**
  - **"Avoid ambiguity."**
  - **"Be brief (avoid unnecessary prolixity)."**
  - **"Be orderly."**

These maxims apply to any cooperative endeavor. Grice uses carpentry as an example. If I am helping you repair your staircase, and you ask for a hammer because you need to drive nails, then:

- **I should hand you one and only one hammer. (Quantity)**
- **I should hand you a non-broken real hammer, as opposed to a broken or toy hammer. (Quality)**
- **I should hand you a carpentry hammer, not a sledge hammer, rubber mallet, or hammer drill. (Relevance)**
- **I should hand you the hammer quickly, dexterously and carefully. (Manner)**

Following are implications for data analysis.

Epistemological Minute

- The Maxim of Quantity: “Make your contribution as informative as is required (for the current purposes of the exchange). Do not make your contribution more informative than is required.”

An edict for data analysts: Reveal as much (but only as much) technical detail as your audience requires. If your audience is researchers, reveal more technical detail. If your audience is practitioners or policymakers, reveal less technical detail. Too much or too little detail will only cause confusion.

- The Maxim of Quality: “Try to make your contribution one that is true.”
  - “Do not say what you believe to be false.”

An edict for data analysts: Be truthful. Truthfulness is a necessary condition for effective data-analytic communication, but as the other maxims imply, it is not a sufficient condition. Truthfulness alone is not helpfulness. “There are lies, damned lies and statistics” (attributed to Disraeli). Statisticians lie, even when they tell the truth, by manipulating/disobeying/distorting the other maxims and playing off expectations of cooperation.

  - “Do not say that for which you lack adequate evidence.”

An edict for data analysts: Be sensitive to your audiences’ standards of evidential adequacy. In statistics, “alpha = .05” sets a standard for adequate evidence to reject the null hypothesis. Consider whether your audience shares that standard. Researchers probably share the standard, but practitioners and policymakers may not.

- The Maxim of Relation: “Be relevant.”

An edict for data analysts: Use statistics logically, not rhetorically. The right statistic at the right time in the conversation can be enlightening if it is logically appropriate. However, that same statistic can be deceiving and stultifying if it is irrelevant. Rarely will statistics be perfectly relevant, so it imperative that data analysts clarify the limitations. For example, most debates are about causes but most statistics are about correlations; the data analyst must make clear when her statistics address the debate only obliquely.

- The Maxim of Manner: “Be perspicuous.”

An edict for data analysts: Write well. Present well. Particularly, do not bury important information, even when it runs contrary to your personal convictions.

Unit 9: Pedagogical Strategy

Everything in this unit is but a short extension from what we have learned in Units 1 through 8. The details, however, will be overwhelming if you give them too much attention. Up until now in the course, I have asked you to keep the pedal to the metal, but now it's time to ease off the gas. I'm not telling you to hit the brakes. I am telling you to coast a little. Ride out your hard fought knowledge.

Part I: Continuous Outcome and Dichotomous Predictor

1. Regression Perspective: 100% Review
2. T-Test Perspective: Same Output Numbers, Different Output Format (with a nice fix for heteroscedasticity)
3. ANOVA Perspective: Same Output Numbers, Different Output Format (with a review of the R^2 statistic)

Part II: Continuous Outcome and Polychotomous Predictor

1. Regression Perspective: Turn the polychotomy into dichotomies. Include all the dichotomies (less one) in our model.
2. ANOVA Perspective: Basic ANOVA output only tells us whether the R^2 statistic is statistically significant, based on the F statistic and the associated p-value. This information is also standard in regression output. However, there are three types of ANOVA supplements that allow us to dig deeper, sometimes a little deeper than regression:
   - A. Contrasts
   - B. Post Hoc Tests
   - C. Plots

		Single Predictor		
		Continuous	Polychotomous	Dichotomous
Outcome	Continuous	Regression	Regression ANOVA	Regression ANOVA T-tests
	Polychotomous	Logistic Regression	Chi Squares	Chi Squares
	Dichotomous		Chi Squares	Chi Squares

Unit 9: Research Question I

Regression Perspective: 100% Review

Theory: Because Anglo students compared with Latino students tend to have greater access to educational resources, Anglo students will tend to perform better than Latino students on tests of academic achievement. This is true even for students who are bound for four-year colleges.

Research Question: In the population of U.S. four-year-college bound boys, do Anglo students, on average, perform better than Latino students on the reading achievement test?

Data Set: NELSBoys.sav National Education Longitudinal Survey (1988), a subsample of 1820 four-year-college bound boys, of whom 182 are Latino and the rest are Anglo.

Variables:

Outcome—Reading Achievement Score (*READ*)

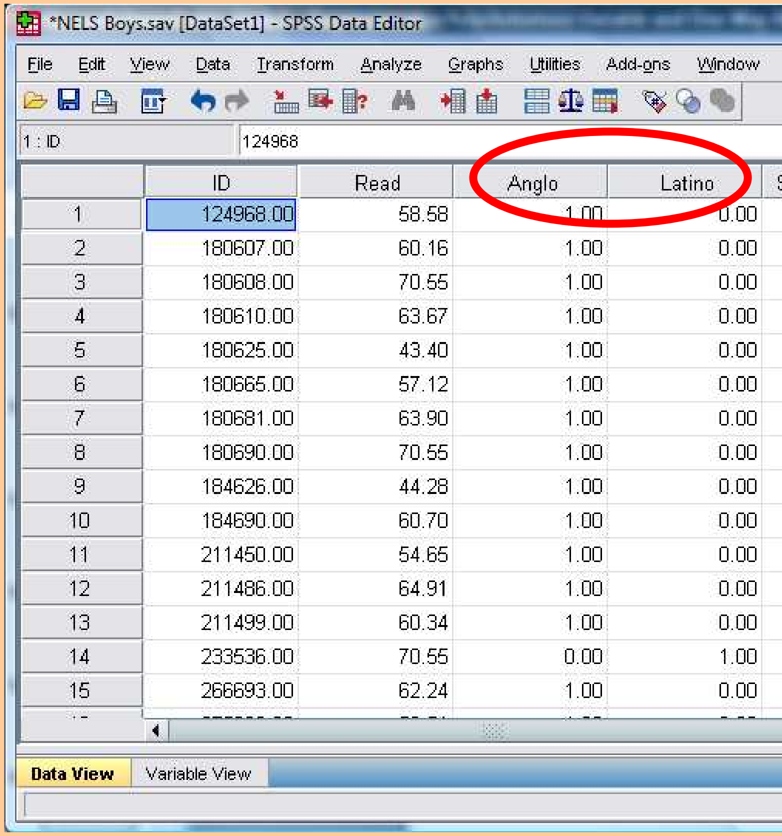
Predictor—Latino = 1, Anglo = 0 (*LATINO*)

Model: $READ = \beta_0 + \beta_1 LATINO + \varepsilon$

Where are Anglos in our model? They are in there. They are the <u>reference category</u>.

NELSBoys.sav

Regression Perspective: 100% Review



	ID	Read	Anglo	Latino
1	124968.00	58.58	1.00	0.00
2	180607.00	60.16	1.00	0.00
3	180608.00	70.55	1.00	0.00
4	180610.00	63.67	1.00	0.00
5	180625.00	43.40	1.00	0.00
6	180665.00	57.12	1.00	0.00
7	180681.00	63.90	1.00	0.00
8	180690.00	70.55	1.00	0.00
9	184626.00	44.28	1.00	0.00
10	184690.00	60.70	1.00	0.00
11	211450.00	54.65	1.00	0.00
12	211486.00	64.91	1.00	0.00
13	211499.00	60.34	1.00	0.00
14	233536.00	70.55	0.00	1.00
15	266693.00	62.24	1.00	0.00

Earlier in the course (Unit 6), we considered degrees of freedom with regard to subjects. For example, when we calculate the mean, we have degrees of freedom equal to the sample size, n (where, n = # of subjects); when we calculate the standard deviation, however, we only have n-1 degrees of freedom (df = n-1), because we use the mean in our calculation of standard deviation. If you tell me the mean value, and you begin listing the values for each subject, then I can finish your list by telling you the value for the last observation. Your last subject provides no unique information for calculating the standard deviation!

In addition to degrees of freedom for subjects, we are going to begin considering degrees of freedom for variables. *ANGLO* and *LATINO* are two variables (k = # of variables = 2), but between them, they only contribute one variable's worth of information, i.e., they contribute one degree of freedom (df = k-1 = 1). If you know one variable, then you know the other.

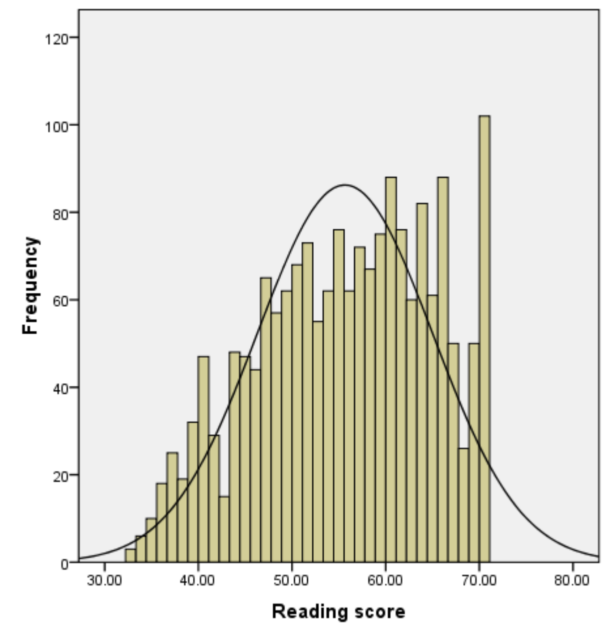
Notice that the information contained in these two variables (*ANGLO* and *LATINO*) is perfectly redundant. If you know one, you know the other. This allows us to leave one out of the model. Our choice to leave out ANGLO, and consequently make it our <u>reference category</u>, was arbitrary. It will have zero impact on our findings, but it will have important consequences for our interpretation, because the parameter estimate for the y-intercept tells us the average for our reference category (when we code 0/1), so we better remember which category is our reference category. Recall that the y-intercept tells us our predicted outcome (e.g., *READ*) when our predictor equals zero (e.g., *LATINO* = 0, which means the student is Anglo).

Exploratory Data Analysis

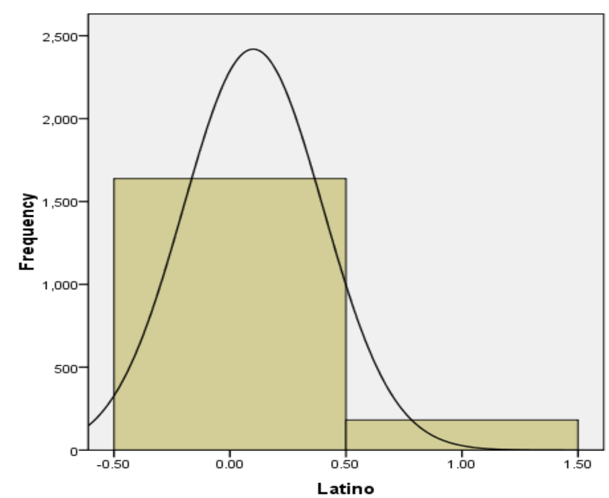
Regression Perspective: 100% Review

Statistics

		Reading score	Latino
N	Valid	1820	1820
	Missing	0	0
Mean		55.6474	.1000
Std. Deviation		9.35512	.30008
Percentiles	25	48.6250	.0000
	50	56.4350	.0000
	75	63.3550	.0000



The median of *READ* is 56.4, which tells us that 50% of our sample scored under 56 points on the reading test. Note the ceiling effect as many students attained the maximum of 70 points.



The mean of *LATINO* is 0.10, which tells us that 10% of our sample self identifies as Latino. This is a nifty interpretation of the mean when we code dichotomous variables with zeroes and ones.

Regression (SPSS)

Regression Perspective: 100% Review

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER Latino.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.161[a]	.026	.025	9.23559

a. Predictors: (Constant), Latino

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4127.193	1	4127.193	48.387	.000[a]
	Residual	155068.466	1818	85.296		
	Total	159195.659	1819			

a. Predictors: (Constant), Latino

b. Dependent Variable: Reading score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	56.149	.228		246.058	.000	55.702	56.597
	Latino	-5.020	.722	-.161	-6.956	.000	-6.435	-3.604

a. Dependent Variable: Reading score

Regression (SPSS)

Regression Perspective: 100% Review

Notice that our estimate for Latinos is not as precise as our estimate for Anglos, as evidenced by the confidence intervals (and the standard errors).

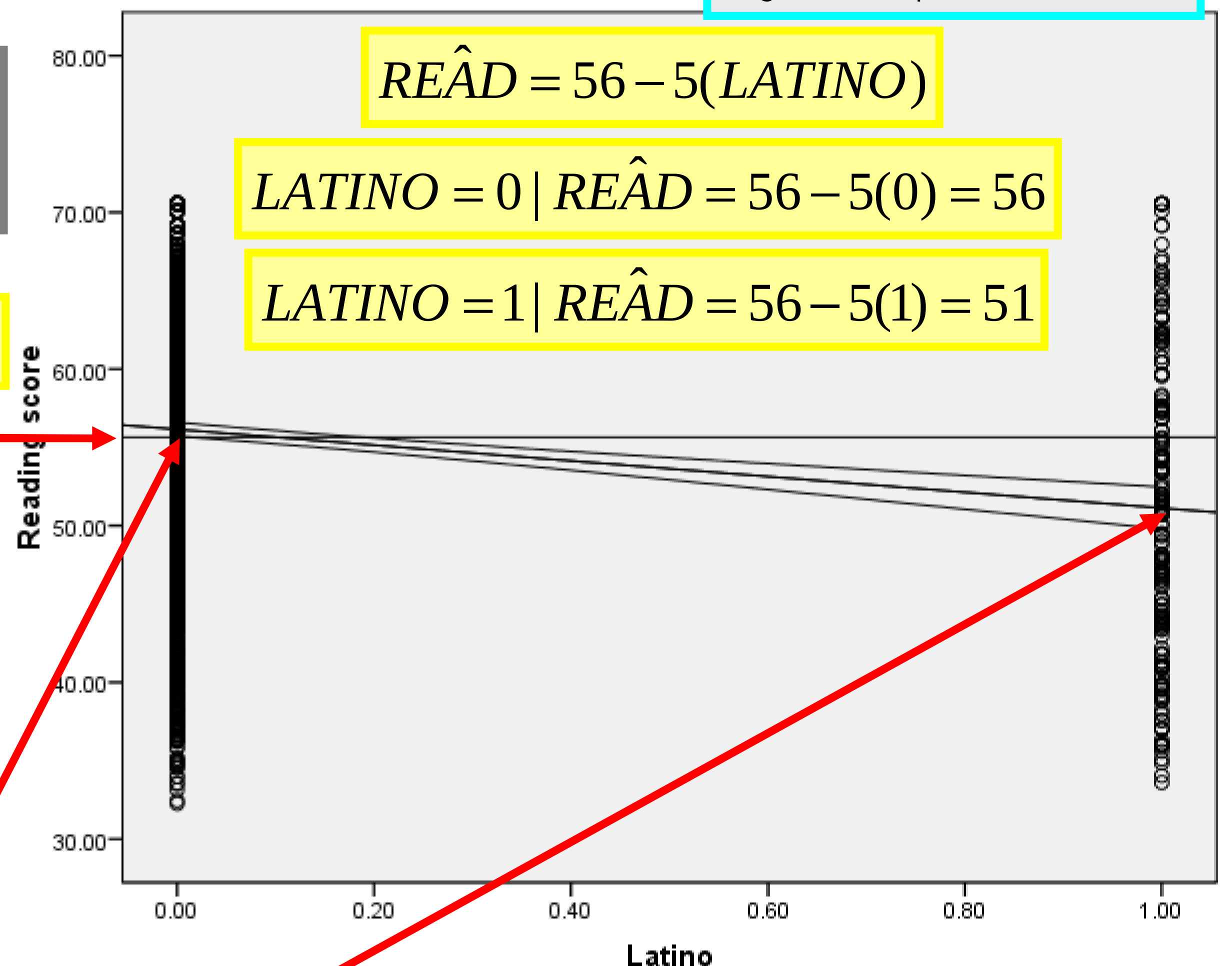
$$READ = \beta_0 + \beta_1 LATINO + \varepsilon$$

$$\hat{READ} = 56 - 5(LATINO)$$

$$LATINO = 0 \,|\, \hat{READ} = 56 - 5(0) = 56$$

$$LATINO = 1 \,|\, \hat{READ} = 56 - 5(1) = 51$$

Grand Mean



The difference that we observe in our sample, five points, is statistically significant (p < 0.001). We estimate that the Latino/Anglo reading gap is between 6.5 and 3.5 points in the population of four-year-college bound boys. We emphasize that we are predicting group averages, not individuals. The best Latino reader in our sample reads as well as the best Anglo reader, and the worst Latino reader in our sample reads better than the worst Anglo reader.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	56.149	.228		246.058	.000	55.702	56.597
	Latino	-5.020	.722	-.161	-6.956	.000	-6.435	-3.604

a. Dependent Variable: Reading score

Regression (R)

Regression Perspective: 100% Review

```
load("E:/Datasets/NELSBoys/nelsboys.rda")
# you can get the regression output in one line
summary(lm(nelsboys$read~nelsboys$latino))
# or you can do it in two lines by naming your model first
my.model <- lm(nelsboys$read~nelsboys$latino)
summary(my.model)
# above is one way to specify the data set. here is a second
my.model <- lm(read~latino, data=nelsboys)
summary(my.model)
# here is a third
attach(nelsboys)
my.model <- lm(read~latino)
summary(my.model)
detach(nelsboys)
```

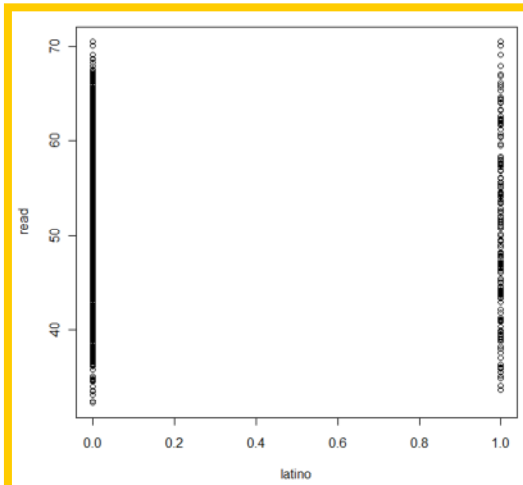
```
Call:
lm(formula = read ~ latino)

Residuals:
     Min       1Q   Median       3Q      Max
-23.8793  -6.8493   0.7457   7.4857  19.4203

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.1493     0.2282 246.058  < 2e-16 ***
latino       -5.0196     0.7216  -6.956 4.86e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.236 on 1818 degrees of freedom
Multiple R-squared: 0.02593,   Adjusted R-squared: 0.02539
F-statistic: 48.39 on 1 and 1818 DF,  p-value: 4.865e-12
```

```
# two ways to plot
# we'll identify the datset by
# attaching it, but either of
# the other methods will work!
attach(nelsboys)
# first method, specify model
plot(read~latino)
# second method, specify x & y
plot(latino, read)
# now, we'll detach the data
detach(nelsboys)
```



At its simplest, R is very simple!

But, R provides the flexibility to ratchet up the complexity and functionality to your heart's desire.

And, it's free!

Two-Sample T-Test (SPSS)

T-Test Perspective: Same Output Numbers, Different Output Format (with a nice fix for heteroscedasticity)

Group Statistics

	Latino	N	Mean	Std. Deviation	Std. Error Mean
Reading score	0	1638	56.1493	9.17302	.22665
	1	182	51.1297	9.78339	.72519

```
T-TEST GROUPS=Latino(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=Read
  /CRITERIA=CI(.9500).
```

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Reading score	Equal variances assumed	1.747	.186	6.956	1818	.000	5.01962	.72162	3.60433	6.43490
	Equal variances not assumed			6.607	217.857	.000	5.01962	.75979	3.52214	6.51709

If our population does not meet the homoscedasticity assumption, then we can use this row: "Equal variances not assumed." Yippy skippy! Levene's Test for Equality of Variances might be informative here. The null hypothesis for the test is that the variances are equal (i.e., the data are homoscedastic). If, based on the p-value being less then 0.05, we reject the null, then we conclude that the data are heteroscedastic, and we use the nifty second line.

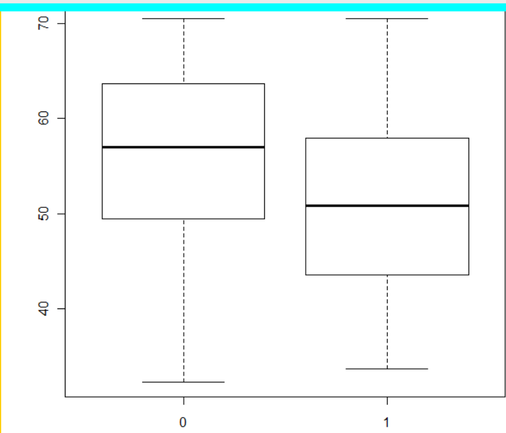
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	56.149	.228		246.058	.000	55.702	56.597
	Latino	-5.020	.722	-.161	-6.956	.000	-6.435	-3.604

a. Dependent Variable: Reading score

Two-Sample T-Test (R)

T-Test Perspective: Same Numbers, Different Format

```
load("E:/Datasets/NELSBoys/nelsboys.rda")
# let’s attach the data, so we don’t have to specify it repeatedly
attach(nelsboys)
# here is the script for a t-test assuming homoscedasticity
t.test(read~latino, var.equal=TRUE)
# here is the script for a t-test allowing for heteroscedasticity
t.test(read~latino, var.equal=FALSE)
# by default, heteroscedasticity is allowed!
t.test(read~latino)
# while we’re at it, let’s examine boxplots of reading by latino
boxplot(read~latino)
# detach the data now that we are done with it
detach(nelsboys)
```




```
   Two Sample t-test
data:  read by latino
t = 6.956, df = 1818, p-value = 4.865e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.604326 6.434905
sample estimates:
mean in group 0 mean in group 1
       56.14934        51.12973
```

Homoscedasticity Assumed!

```
   Welch Two Sample t-test
data:  read by latino
t = 6.6066, df = 217.857, p-value = 2.971e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.522143 6.517087
sample estimates:
mean in group 0 mean in group 1
       56.14934        51.12973
```

Heteroscedasticity Allowed!

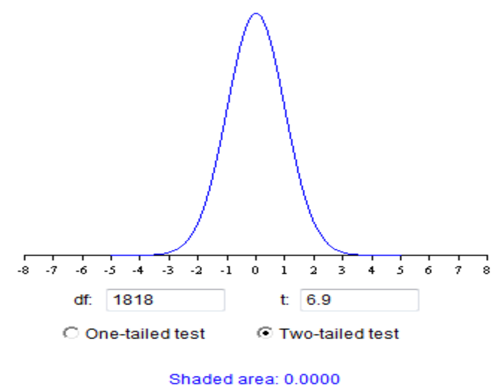
Two-Sample T-Test (By Hand)

T-Test Perspective: Same Numbers, Different Format

Group Statistics

	Latin 0	N	Mean	Std. Deviation	Std. Error Mean
Reading score	0	1638	56.1493	9.17302	.22665
	1	182	51.1297	9.78339	.72519

http://onlinestatbook.com/analysis_lab/t_dist.html



$$t = \frac{\text{difference in means}}{\text{standard error}_{\text{mean}_1\text{-mean}_2}} = \frac{5}{0.72} = 6.9$$

$$\text{standard error}_{\text{mean}_1\text{-mean}_2} = \sqrt{\frac{(1638-1)9.17^2 + (182-1)9.78^2}{(1638-1)+(182-1)}} \times \sqrt{\frac{1}{1638}+\frac{1}{182}} = 0.72$$

Notice that, aside from a little squaring, a little rooting, and a 1 here and there, the only numbers are the standard deviations and the sample sizes. Standard errors, at their core, stem from the Central Limit Theorem which says that the standard deviation of a sampling distribution of means is the population standard deviation (σ) divided by the square root of the sample size (√n).

$$\text{standard error}_{\text{mean}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

One-Way ANOVA (SPSS)

ANOVA Perspective: Same Output Numbers, Different Output Format (with a review of the R^2 statistic)

Tests of Between-Subjects Effects

Dependent Variable:Reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4127.193[a]	1	4127.193	48.387	.000
Intercept	1885141.110	1	1885141.110	22101.119	.000
Latino	4127.193	1	4127.193	48.387	.000
Error	155068.466	1818	85.296		
Total	5795063.720	1820			
Corrected Total	159195.659	1819			

a. R Squared = .026 (Adjusted R Squared = .025)

```
UNIANOVA Read BY Latino
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /CRITERIA=ALPHA(0.05)
  /DESIGN=Latino.
```

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4127.193	1	4127.193	48.387	.000[a]
	Residual	155068.466	1818	85.296		
	Total	159195.659	1819			

Notice the t statistic for the slope: -6.956.
If we square it, we get the F statistic: 48.387.
This trick works when we have one variable in the model. We'll introduce the F statistic today and go into more detail in Unit 10.

***Recall from Unit 5 that the Total Sum of Squares is the sum of squared deviations from the grand mean, in this case, the mean of *READ*. The total sum of squares sets our baseline for the variation that needs predicting.**
***The Residual (or Error) Sum of Squares is the sum of squared deviations from our regression line. Once we make our prediction(s), we want the sum of squared deviations to be small. Small compared to what? Small compared to our baseline, Total Sum of Squares.**
***Here, our Residual (or Error) Sum of Squares is 0.976 of the Total Sum of Squares. Therefore, our R^2 statistic is 0.026. Is the R^2 statistic of 0.026 stat sig? The omnibus F test tells us so, $F(1,1818)=48.39$, $p<0.001$. Think of the omnibus F test as testing the null hypothesis that the population R^2 is zero.**

One-Way ANOVA (R)

ANOVA Perspective: Same Output Numbers, Different Output Format (with a review of the R^2 statistic)

```
load("E:/Datasets/NELSBoys/nelsboys.rda")
# attach the dataset
attach(nelsboys)
# name your model
model.1 <- lm(read~latino)
# produce an ANOVA table for your model
anova(model.1)
# detach the dataset
detach(nelsboys)
```

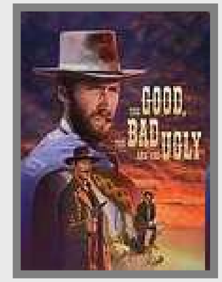
```
attach(nelsboys)
# use R as a fancy calculator
4127+155068
# or compute the variance
var(read)
# and compute the sample size
length(read)
# use that info to get the SST
var(read)*(length(read)-1)
# or just do Post Hole 3!
sum((read-mean(read))^2)
detach(nelsboys)
```

```
Analysis of Variance Table

Response: read
            Df Sum Sq Mean Sq F value    Pr(>F)
latino       1   4127  4127.2  48.387 4.865e-12 ***
Residuals 1818 155068    85.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

This basic R output leaves out the total sum of squares. We can get the total sum of squares in a couple of ways. We can simply add the model sum of squares and the residual sum of squares (4127+155068), or we can multiply the variance of the outcome by the degrees of freedom (n-1) thereby reversing the step in Post Hole 3 from the total sum of squares to the variance.

The Good, The Bad, And The Baseline



For predictive purposes, some variation is good, other variation is bad, and still other variation is neither good nor bad, but baseline. Standard deviation is a measure of variation. On your way to calculating the standard deviation for Post Hole 3, you calculate variance (another measure of variation). On your way to calculating variance, you calculate the sum of squares (yet another measure of variation). We'll focus on the sum of squares as a measure of variation.

<u>Baseline</u> Sums of Squares: Total (or Corrected Total)

The mean is always there for us. If we have an outcome, we can calculate its mean. We don't need no stinkin' predictors! We can treat the mean as our best prediction in the absence of further information. As such, the mean provides an ever-handy basis of comparison for any more informed predictions that we may make.

<u>Bad</u> Sums of Squares: Residual (or Error)

I don't mean "bad" as in Satanic. Rather, from a predictive perspective, where our goal is prediction, it is "bad" when our prediction is wrong. Honestly, sometimes I want my predictions to be wrong; for example, I predict a blizzard. However, from a predictive perspective, if I make a prediction and it's wrong, then that's bad. If the meteorologist forecasts a blizzard and we get a flurry instead of a storm, then that's a strike against the forecaster. Residuals (or errors) measure for each observation how wrong our prediction is. When we square the residuals/errors and sum them, that measure of variation is capturing the badness of fit of our model. Thus, I will call it "bad variation."

<u>Good</u> Sums of Squares: Regression (or Corrected Model)

From a predictive perspective, it is good when our predictions diverge from the mean. After all, if our predictions are no different from the mean, then why bother with the rigmarole of fitting a statistical model? When our predictions are different from the mean, our regression model is adding predictive value over and above the mean, where the mean is just a generic (i.e., unconditional) prediction. The regression/model sum of squares measures the (squared) difference between each prediction and the mean, so I call it "good variation."

Also: Sample size is <u>good</u>; the more, the better. All else being equal, larger samples contribute to more statistically powerful (i.e., more precise) analyses. Predictor variables are (a little) <u>bad</u>; the fewer, the better. Consider two models that provide equally good predictions. The model with the fewer variables is better (all else being equal).

R^2 Statistic Review

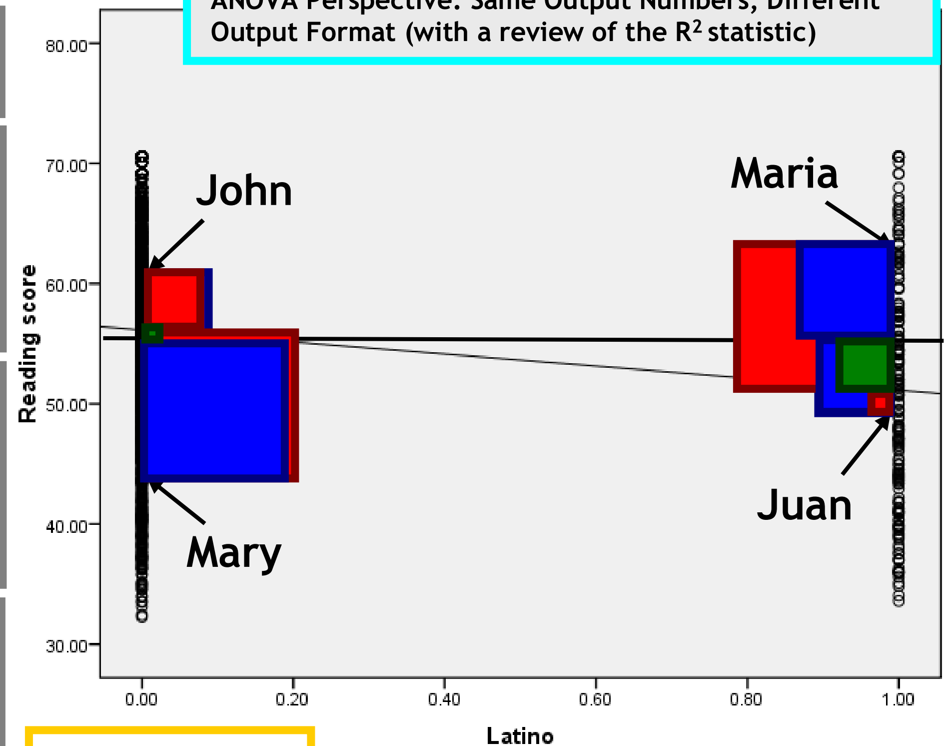
If our goal is prediction, then some variation is good, some variation is bad, and some variation is baseline.

The baseline variation can be measured by summing the squared differences of each observation from the grand mean. It is called the "Total Sum of Squares." In fact, this is simply the sum of squared mean deviations that you calculate for Post Hole 3!

The bad variation can be measured by summing the squared differences of each observation from the regression line, i.e., prediction, i.e., group mean. It is called the "Residual Sum of Squares" or "Error Sum of Squares."

The good variation can be measured by summing the squared differences of the grand mean from the regression line, i.e., prediction, i.e., group mean. It is called the "Regression Sum of Squares" or "Model Sum of Squares."

ANOVA Perspective: Same Output Numbers, Different Output Format (with a review of the R^2 statistic)




```
UNIANOVA Read BY Latino
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=ETASQ
  /CRITERIA=ALPHA(.05)
  /DESIGN=Latino.
```

Tests of Between-Subjects Effects

Dependent Variable:Reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	4127.193[a]	1	4127.193	48.387	.000	.026
Intercept	1885141.110	1	1885141.110	22101.119	.000	.924
Latino	4127.193	1	4127.193	48.387	.000	.026
Error	155068.466	1818	85.296			
Total	5795063.720	1820				
Corrected Total	159195.659	1819				

a. R Squared = .026 (Adjusted R Squared = .025)

The R^2 statistic is the proportion of good variation to baseline variation. Or, since the good variation plus the bad variation equal the baseline variation, the R^2 statistic is one minus the proportion of bad variation to baseline variation.

In ANOVA, the R^2 statistic is called the "Eta squared statistic."

Is the R^2 Statistic (i.e., η^2 Statistic) Statistically Significant?

If our goal is prediction, we want the good variation to outweigh the bad variation. That is an uphill battle! But, we have help! We get to divide the bad variation by the degrees of freedom of the sample (n-2). To be fair, however, we also have to divide our good variation by the degrees of freedom of the variables, but that will be small unless we include a bunch of garbage variables in our model.

Once we divide, we get mean squares. Consequently, there is a good mean square and bad mean square.

Then, we can divide the good mean square by the bad mean square to get the F-statistic, which is good because we want it to be big.

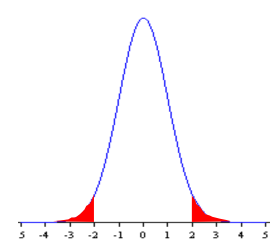
The Central Limit Theorem tells us that shape of the sampling distribution of t-statistics approaches normality as sample size approaches infinity. Similarly, the Central Limit Theorem tells us that the shape of the sampling distribution of F-statistics is positively skewed. We can use the F-distribution to reject the null hypothesis that the R^2 statistic is 0.00000000 in the population.

ANOVA Perspective: Same Output Numbers, Different Output Format (with a review of the R^2 statistic)

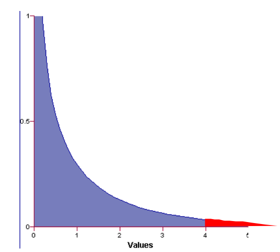
Oops! I lied. ***This*** is *new*, but it's really just a preview of Part II. Please forgive me.

When we have one variable in our model, the F-statistic is simply the square of the t-statistic, and the F-distribution is the square of the t distribution.

http://www.capdm.com/demos/software/html/capdm/qm/fdist/usage.html



t distribution



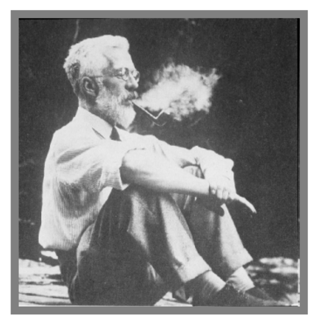
F distribution

As you would expect from the squared relationship, since a t-statistic of about ±2 marks the reject-the-null region, an F-statistic of about 4 marks the reject-the-null region.

Dependent Variable:Reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	4127.193a	1	4127.193	48.387	.000	.026
Intercept	1885141.110	1	1885141.110	22101.119	.000	.924
Latino	4127.193	1	4127.193	48.387	.000	.026
Error	155068.466	1818	85.296			
Total	5795063.720	1820				
Corrected Total	159195.659	1819				

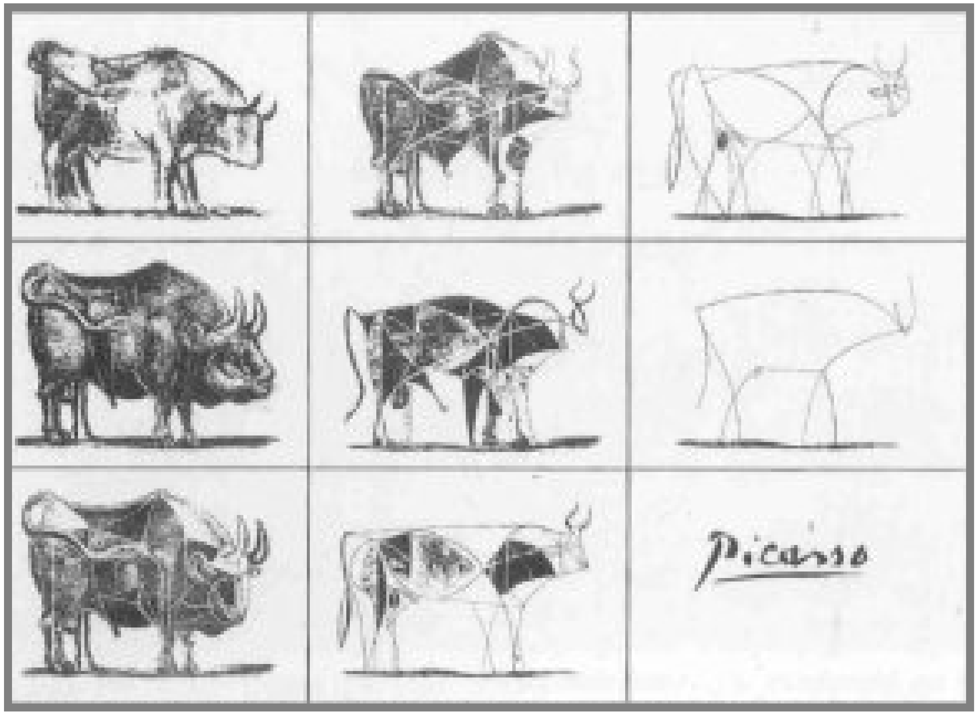
a. R Squared = .026 (Adjusted R Squared = .025)



F is for R.A. Fisher

C is for Cookie

5-Minute Break



Unit 9: Research Question II

Regression Perspective: Turn the polychotomy into dichotomies. Include all the dichotomies (less one) in your model.

Theory: Because higher SES students tend to have greater access to educational resources, high SES students will tend to perform better than mid SES students, and mid SES students better than low SES students, on tests of academic achievement. This is true even for students who are bound for four-year colleges.

Research Question: In the population of U.S. four-year-college bound boys, do higher SES students, on average, perform better than lower SES students on the reading achievement test?

Data Set: NELSBoys.sav National Education Longitudinal Survey (1988), a subsample of 1820 four-year-college bound boys, of whom 182 are Latino and the rest are Anglo.

Variables:

Outcome—Reading Achievement Score (*READ*)

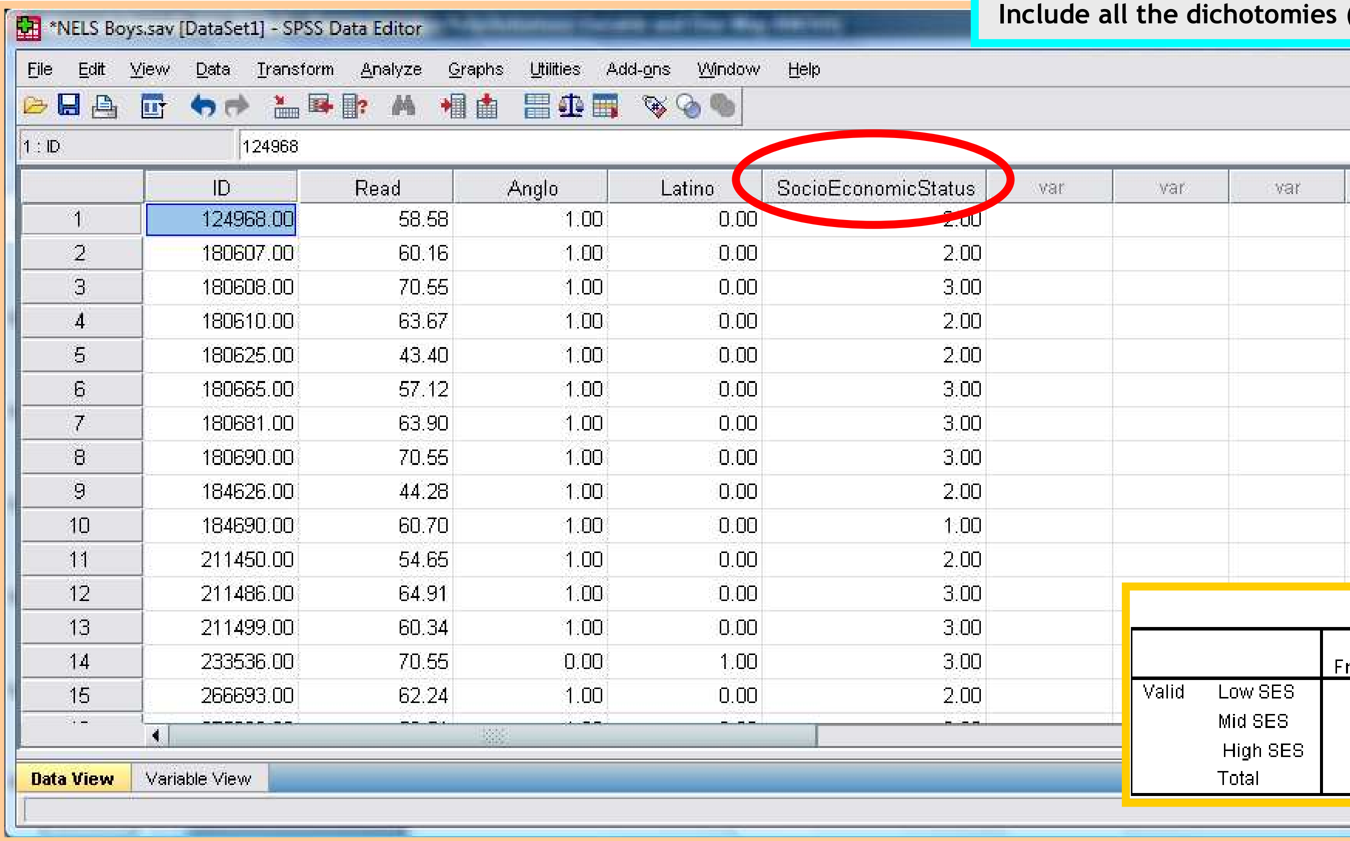
Predictor—Low SES=1, Mid SES=2, High SES=3 (*SocioEconomicStatus*)

Model: $READ = \beta_0 + \beta_1 LowSES + \beta_2 HighSES + \varepsilon$

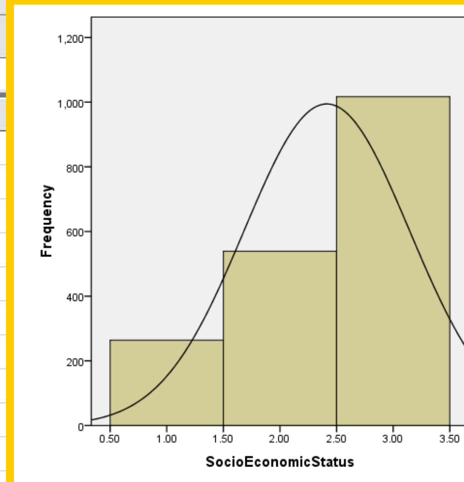
Where is MidSES in our model? It is in there. It is the reference category.

NELSBoys.sav

Regression Perspective: Turn the polychotomy into dichotomies. Include all the dichotomies (less one) in your model.



	ID	Read	Anglo	Latino	SocioEconomicStatus	var	var	var
1	124968.00	58.58	1.00	0.00	2.00			
2	180607.00	60.16	1.00	0.00	2.00			
3	180608.00	70.55	1.00	0.00	3.00			
4	180610.00	63.67	1.00	0.00	2.00			
5	180625.00	43.40	1.00	0.00	2.00			
6	180665.00	57.12	1.00	0.00	3.00			
7	180681.00	63.90	1.00	0.00	3.00			
8	180690.00	70.55	1.00	0.00	3.00			
9	184626.00	44.28	1.00	0.00	2.00			
10	184690.00	60.70	1.00	0.00	1.00			
11	211450.00	54.65	1.00	0.00	2.00			
12	211486.00	64.91	1.00	0.00	3.00			
13	211499.00	60.34	1.00	0.00	3.00			
14	233536.00	70.55	0.00	1.00	3.00			
15	266693.00	62.24	1.00	0.00	2.00			



SocioEconomicStatus

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Low SES	264	14.5	14.5	14.5
	Mid SES	539	29.6	29.6	44.1
	High SES	1017	55.9	55.9	100.0
	Total	1820	100.0	100.0	

Here, *SocioEconomicStatus* is a categorical variable, with three categories. We do not trust that the scale is interval. In other words, we are not convinced that the difference between a 1 and a 2 is the same as the difference between a 2 and 3. In reality, we have this doubt about most scales in the social sciences. The question is not whether our scale is interval or not. Rather, the question is whether our scale is interval enough. When we look at the distribution of *SocioEconomicStatus*, it is doubtful that the scale is interval enough, so we will treat the three categories as ordinal.

NELSBoys.sav

Regression Perspective: Turn the polychotomy into dichotomies. Include all the dichotomies (less one) in your model.

*NELS Boys.sav [DataSet1] - SPSS Data Editor

	ID	Read	Anglo	Latino	SocioEconomicStatus	LowSES	HighSES
1	124968.00	58.58	1.00	0.00	2.00	0.00	0.00
2	180607.00	60.16	1.00	0.00	2.00	0.00	0.00
3	180608.00	70.55	1.00	0.00	3.00	0.00	1.00
4	180610.00	63.67	1.00	0.00	2.00	0.00	0.00
5	180625.00	43.40	1.00	0.00	2.00	0.00	0.00
6	180665.00	57.12	1.00	0.00	3.00	0.00	1.00
7	180681.00	63.90	1.00	0.00	3.00	0.00	1.00
8	180690.00	70.55	1.00	0.00	3.00	0.00	1.00
9	184626.00	44.28	1.00	0.00	2.00	0.00	0.00
10	184690.00	60.70	1.00	0.00	1.00	1.00	0.00
11	211450.00	54.65	1.00	0.00	2.00	0.00	0.00
12	211486.00	64.91	1.00	0.00	3.00	0.00	1.00
13	211499.00	60.34	1.00	0.00	3.00	0.00	1.00
14	233536.00	70.55	0.00	1.00	3.00	0.00	1.00
15	266693.00	62.24	1.00	0.00	2.00	0.00	0.00

Although *SocioEconomicStatus* has three categories (k = 3), and we could thus create three dichotomies, they would only have two degrees of freedom (df = k -1 = 2). So, we could create a variable, *MidSES*, but it would provide no information over and above the information provided by *LowSES* and *HighSES*. A student for whom *MidSES* would equal one is a student for whom *LowSES* and *HighSES* both equal zero.

Since we know how to deal with dichotomous predictors, we'll turn our polychotomous predictor into a set of (0/1) dichotomous predictors (aka "dummy variables" or "indicator variables").

```
Compute LowSES  = 0.
If (SocioEconomicStatus = 1) LowSES = 1.
Compute HighSES  = 0.
If (SocioEconomicStatus = 3) HighSES = 1.
Execute.
```

Regression (SPSS)

Regression Perspective: Turn the polychotomy into dichotomies. Include all the dichotomies (less one) in your model.

MidSES is the <u>reference category</u>. The y-intercept represents the average reading score for middle SES students.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER LowSES HighSES.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.211[a]	.044	.043	9.15038

a. Predictors: (Constant), HighSES, LowSES

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7059.269	2	3529.634	42.155	.000[a]
	Residual	152136.390	1817	83.729		
	Total	159195.659	1819			

a. Predictors: (Constant), HighSES, LowSES

b. Dependent Variable: Reading score

Coefficients[a]

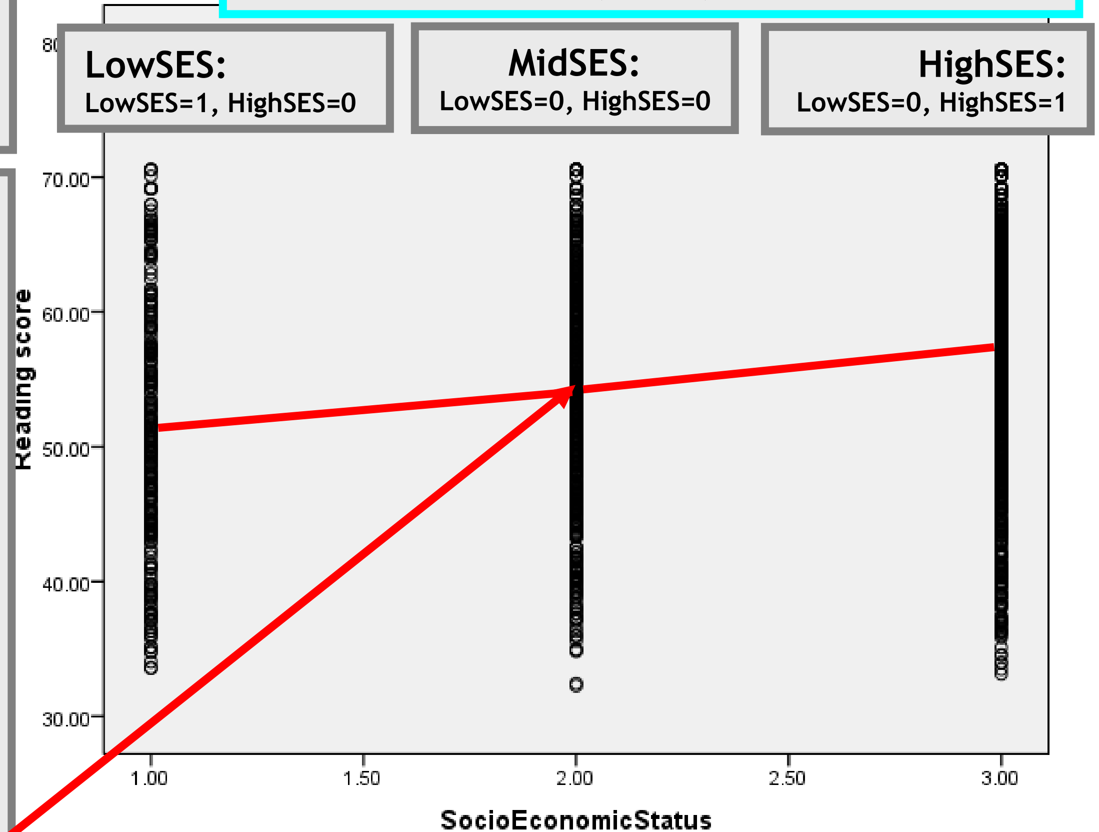
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.418	.394		138.070	.000	53.645	55.191
	LowSES	-2.466	.687	-.093	-3.588	.000	-3.815	-1.118
	HighSES	2.840	.488	.151	5.826	.000	1.884	3.796

a. Dependent Variable: Reading score

Regression (SPSS)

Regression Perspective: Turn the polychotomy into dichotomies. Include all the dichotomies (less one) in your model.

Perhaps one straight line would have been appropriate! (We'll see next week why perhaps not.)



We could have chosen any of the three categories as our reference category by leaving it out. If we chose to leave out LowSES, we would see:

(Constant)	51.953
MidSES	2.466
HighSES	5.306

If we chose HighSES, we would see:

(Constant)	57.218
LowSES	-5.306
MidSES	-2.840

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.418	.394		138.070	.000	53.645	55.191
	LowSES	-2.466	.687	-.093	-3.588	.000	-3.815	-1.118
	HighSES	2.840	.488	.151	5.826	.000	1.884	3.796

a. Dependent Variable: Reading score

Regression on Dummies

Regression Perspective: Turn the polychotomy into dichotomies. Include all the dichotomies (less one) in your model.

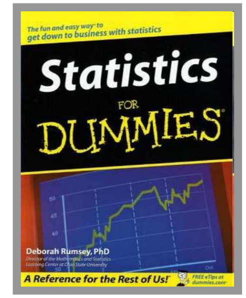
Although technically we are regressing *READ* on multiple variables, there is conceptually only one predictor variable: *SocioEconomicStatus*.

We will save multiple regression for next week, Unit 10.

In regression, we are ultimately trying to predict on average. When we have a polychotomy with three categories, we are using three predictive averages.

$$READ = \beta_0 + \beta_1 LowSES + \beta_2 HighSES + \varepsilon$$

$$\hat{READ} = 54.4 - 2.5(LowSES) + 2.8(HighSES)$$



When SocioEconomicStatus = 1, LowSES = 1 and HighSES = 0.

$$SocioEconomicStatus = 1 \mid \hat{READ} = 54.4 - 2.5(1) + 2.8(0) = 51.9$$

When SocioEconomicStatus = 2, LowSES = 0 and HighSES = 0.

$$SocioEconomicStatus = 2 \mid \hat{READ} = 54.4 - 2.5(0) + 2.8(0) = 54.4$$

When SocioEconomicStatus = 3, LowSES = 0 and HighSES = 1.

$$SocioEconomicStatus = 3 \mid \hat{READ} = 54.4 - 2.5(0) + 2.8(1) = 57.2$$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.418	.394		138.070	.000	53.645	55.191
	LowSES	-2.466	.687	-.093	-3.588	.000	-3.815	-1.118
	HighSES	2.840	.488	.151	5.826	.000	1.884	3.796

a. Dependent Variable: Reading score

The Omnibus F Test

The R^2 statistic is about the proportion of variation predicted by the entire model. Up until know, we have treated it as the proportion of variation predicted by the predictor because up until know we only had one predictor in our model.

Regression Perspective: Turn the polychotomy into dichotomies. Include all the dichotomies (less one) in your model.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER LowSES HighSES.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.211[a]	.044	.043	9.15038

a. Predictors: (Constant), HighSES, LowSES

The omnibus F test tells us whether the R^2 statistic is stat sig.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7059.269	2	3529.634	42.155	.000[a]
	Residual	152136.390	1817	83.729		
	Total	159195.659	1819			

a. Predictors: (Constant), HighSES, LowSES

b. Dependent Variable: Reading score

The t tests tell us whether the individual dummies are stat sig contributors to the R^2 statistic.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.418	.394		138.070	.000	53.645	55.191
	LowSES	-2.466	.687	-.093	-3.588	.000	-3.815	-1.118
	HighSES	2.840	.488	.151	5.826	.000	1.884	3.796

a. Dependent Variable: Reading score

Regression (R)

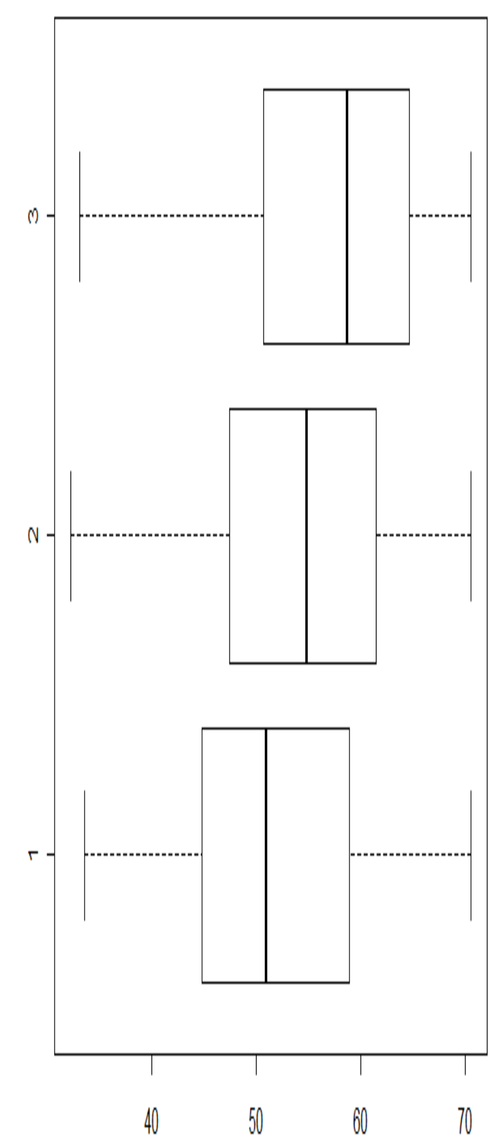
Regression Perspective: Turn the polychotomy into dichotomies. Include all the dichotomies (less one) in your model.

```
load("E:/Datasets/NELSBoys/nelsboys.rda")
attach(nelsboys)
# turn the polychotomy into dichotomies
# “socioeconomicstatus==1” creates a vector of TRUEs and FALSEs as it checks
# each value of socioeconomicstatus to see whether it is equal to 1
# “as.numeric()” coerces the TRUEs and FALSEs to 1s and 0s, respectively
# our now familiar “<-” names (or “assigns”)the results for future use
low.ses <- as.numeric(socioeconomicstatus==1)
mid.ses <- as.numeric(socioeconomicstatus==2)
high.ses <- as.numeric(socioeconomicstatus==3)
# fit the linear model (lm)
model.2 <- lm(read ~ low.ses + high.ses)
summary(model.2)
# here is an R shortcut for polychotomies
# treat socioeconomicstatus as a factor, calling the factor whatever
ses.factor <- as.factor(socioeconomicstatus)
model.3 <- lm(read ~ ses.factor)
summary(model.3)
# let’s create a boxplot for kicks
boxplot(read ~ socioeconomicstatus, horizontal=TRUE)
detach(nelsboys)
```

```
lm(formula = read ~ low.ses + high.ses)
Residuals:
    Min      1Q  Median      3Q     Max
-24.118  -6.723   0.762   7.202  18.598

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.4181     0.3941 138.070  < 2e-16 ***
low.ses      -2.4664     0.6874  -3.588 0.000342 ***
high.ses      2.8402     0.4875   5.826 6.71e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.15 on 1817 degrees of freedom
Multiple R-squared: 0.04434,   Adjusted R-squared: 0.04329
F-statistic: 42.16 on 2 and 1817 DF,  p-value: < 2.2e-16
```



One-Way ANOVA (SPSS)

ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
SocioEconomicStatus	1	Low SES	264
	2	Mid SES	539
	3	High SES	1017

```
UNIANOVA Read BY SocioEconomicStatus
  /CONTRAST(SocioEconomicStatus)=Simple (2)
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=ETASQ
  /POSTHOC=SocioEconomicStatus(BONFERRONI)
  /PLOT=PROFILE(SocioEconomicStatus)
  /CRITERIA=ALPHA(0.05)
  /DESIGN=SocioEconomicStatus.
```

Tests of Between-Subjects Effects

Dependent Variable:Reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	7059.269[a]	2	3529.634	42.155	.000	.044
Intercept	4040493.533	1	4040493.533	48256.546	.000	.964
SocioEconomicStatus	7059.269	2	3529.634	42.155	.000	.044
Error	152136.390	1817	83.729			
Total	5795063.720	1820				
Corrected Total	159195.659	1819				

a. R Squared = .044 (Adjusted R Squared = .043)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7059.269	2	3529.634	42.155	.000[a]
	Residual	152136.390	1817	83.729		
	Total	159195.659	1819			

a. Predictors: (Constant), HighSES, LowSES

b. Dependent Variable: Reading score

With ANOVA, we throw the polychotomy into the hopper all at once, with no need to dichotomize into dummies. However, the basic ANOVA only tells us that there is a relationship. It does not tell us where the relationship is or how big it is except for the R^2 (or η^2) statistic. Nevertheless, we can get that info from contrasts (planned comparisons), post hoc tests (unplanned comparisons), and plots.

One-Way ANOVA (R)

ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.

```
load("E:/Datasets/NELSBoys/nelsboys.r
attach(nelsboys)
ses.factor <- as.factor(socioeconomicstatus)
model.2 <- lm(read ~ ses.factor)
anova(model.3)
detach(nelsboys)
```

```
Analysis of Variance Table

Response: read
             Df Sum Sq Mean Sq F value    Pr(>F)
ses.factor    2   7059  3529.6  42.155 < 2.2e-16 ***
Residuals  1817 152136    83.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

Analysis of Variance (ANOVA): Analyzing What Variance?

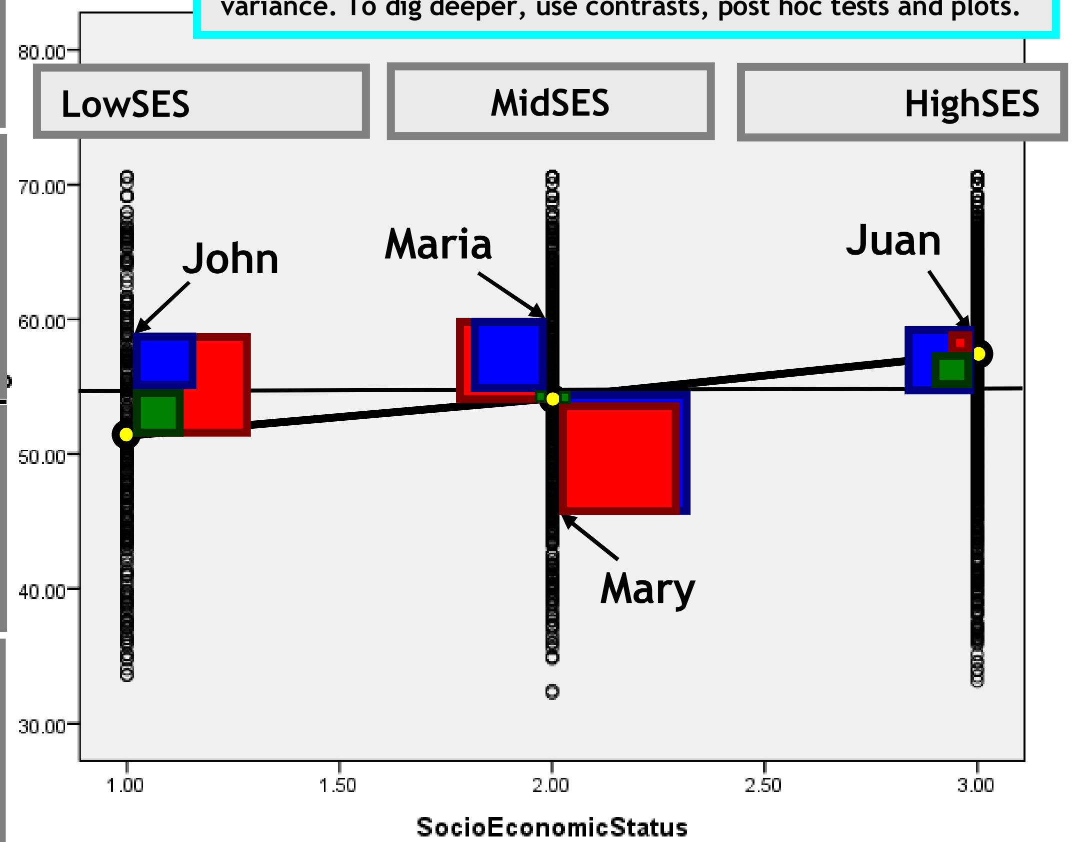
ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.

If our goal is prediction, then some variation is good, some variation is bad, and some variation is baseline.

The baseline variation can be measured by summing the squared differences of each observation from the grand mean. It is called the "Total Sum of Squares." In fact, this is simply the sum of squared mean deviations that you calculate for Post Hole 3!

The bad variation can be measured by summing the squared differences of each observation from the regression line, i.e., prediction, i.e., group mean. It is called the "Residual Sum of Squares" or "Error Sum of Squares."

The good variation can be measured by summing the squared differences of the grand mean from the regression line, i.e., prediction, i.e., group mean. It is called the "Regression Sum of Squares" or "Model Sum of Squares."



The variance that we are analyzing is the variance of the outcome, the baseline variation. From statistical algebra, we know that the good variation plus the bad variation equals the baseline variation. In other words, if we made squares for every observation (not just four), then the green squares plus the red squares would equal the blue squares. Thus, we can partition (i.e., analyze) the baseline variation into good variation and bad variation. We can say that the good variation is 4.4% of the baseline variation ($R^2 = .044$). We can also consider the ratio of good variation to bad variation, $F(2, 1817) = 42.15$, $p < .001$. (We'll explore this further in the next slide.) Some researchers think of the good variation as "signal" and the bad variation as "noise"; thus, the F statistic is the ratio of signal to noise.

The F Statistic: What Is It Good For?

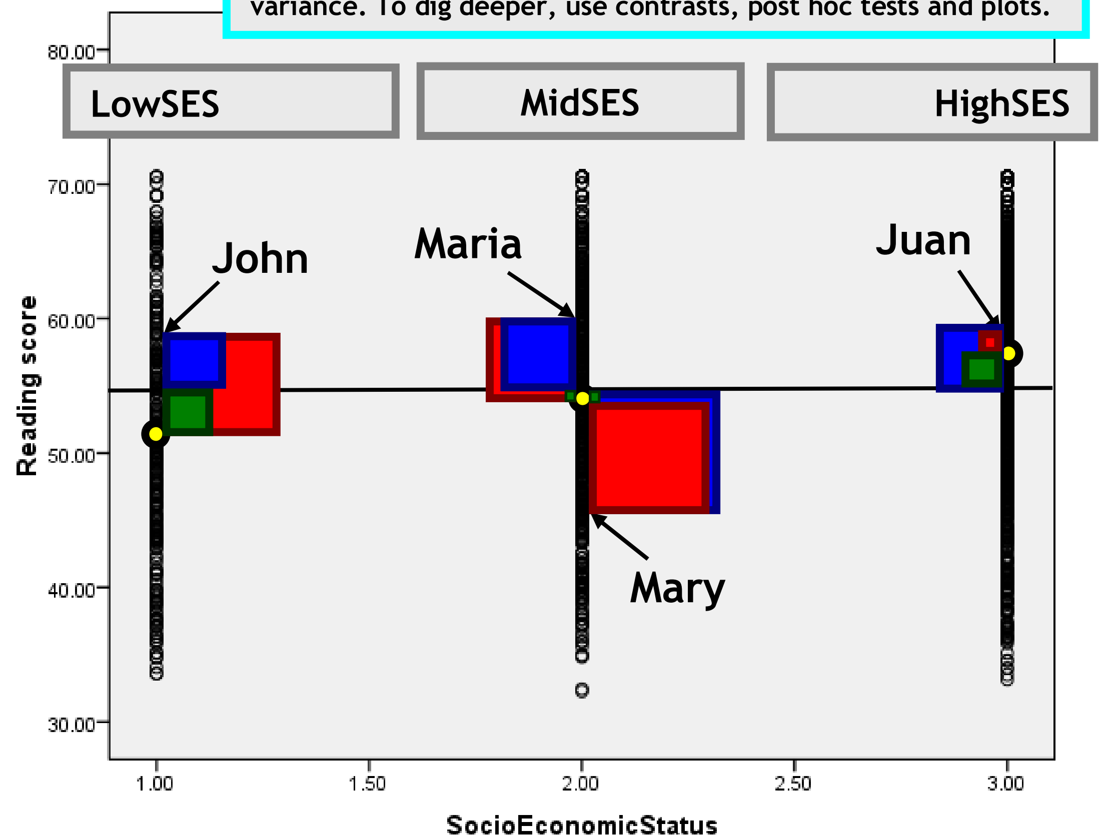
If our goal is prediction, we want the good variation to outweigh the bad variation. That is an uphill battle! But, we have help! We get to divide the bad variation by the degrees of freedom of the sample (e.g., n-2). To be fair, however, we also have to divide our good variation by the degrees of freedom of the variables (e.g., k-1), but that will be small unless we include a bunch of garbage variables in our model.

Once we divide, we get mean squares. Consequently, there is a good mean square and bad mean square.

Then, we can divide the good mean square by the bad mean square to get the F-statistic, which is good because we want it to be big. This is a ratio of signal to noise.

Is the F statistic statistically significant?

ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.



Dependent Variable:Reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	7059.269[a]	2	3529.634	42.155	.000	.044
Intercept	4040493.533	1	4040493.533	48256.546	.000	.964
SocioEconomicStatus	7059.269	2	3529.634	42.155	.000	.044
Error	152136.390	1817	83.729			
Total	5795063.720	1820				
Corrected Total	159195.659	1819				

a. R Squared = .044 (Adjusted R Squared = .043)

I.e., is the good variation too big on average to be plausibly accidental? Is our sample (R^2=.044) plausibly drawn from a population with R^2=0.00. Could the "signal" plausibly be just noise?

Is the F Statistic Statistically Significant?

ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.

Mr. Null

So, in your sample, the ratio of good variation to bad variation is greater than zero? In your sample, your ratio of signal to noise is positive? In your sample, F > 0? I, Mr. Null, hypothesize that, in the population, there is no relationship between your outcome and your combined predictors. I hypothesize that in the population the F statistic and thus the R^2 statistic are zero. Your results are merely due to sampling error.

The key to addressing any null hypothesis is to consider the sampling distribution for your statistic. If you took a thousand (equally sized) random samples from the population, and you calculated your statistic for each sample, how would the statistics distribute themselves? Whereas means and slopes form a normal distribution (or t distribution), F statistics form a positively skewed distribution the exact shape of which depends on not only the degrees of freedom of the subjects but also the degrees of freedom of the variables. Once you have your sampling distribution, you can set it at zero (as per the null hypothesis) and observe if your statistic is far enough away from zero (based on your alpha level) to reject the null hypothesis.

If you stare at TV static for a while, you will start seeing patterns. Sometimes people see departed loved ones. Spooky.

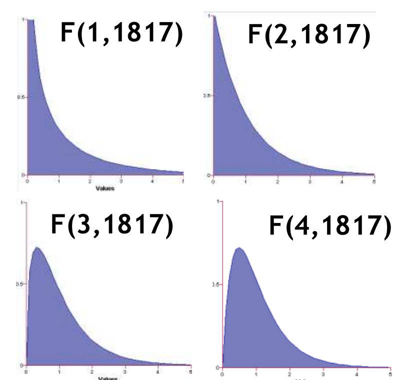
There is a statistically significant relationship between socioeconomic status and reading scores, $F(2,1817) = 42.155$, $p < .001$. The null hypothesis is that there is no relationship in the population. We reject the null hypothesis based on a p-value of less than .05. We conclude that there is a relationship in the population.

Dependent Variable: Reading score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	7059.269[a]	2	3529.634	42.155	.000	.044
Intercept	4040493.533	1	4040493.533	48256.546	.000	.964
SocioEconomicStatus	7059.269	2	3529.634	42.155	.000	.044
Error	152136.390	1817	83.729			
Total	5795063.720	1820				
Corrected Total	159195.659	1819				

a. R Squared = .044 (Adjusted R Squared = .043)

Four F Distributions



Planned Comparisons: Contrasts

ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.

Contrast Results (K Matrix)

SocioEconomicStatus Simple Contrast[a]			Dependent Variable: Reading score
Level 1 vs. Level 2	Contrast Estimate		-2.466
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-2.466
	Std. Error		.687
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-3.815
		Upper Bound	-1.118
Level 3 vs. Level 2	Contrast Estimate		2.840
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		2.840
	Std. Error		.488
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	1.884
		Upper Bound	3.796

a. Reference category = 2

There are many planned comparisons (i.e., contrasts), but I would like to introduce you to these three for starters. We can replicate them all in regression, but it involves tricks of coding that are better left for future semesters. The key is to judiciously sprinkle -1s (and other numbers) into your 0/1 coding. Right now, you can do "Simple" Contrasts in regression.

The relationship is stat sig. Now what?

Simple Contrasts (above) compare every level of your factor to the level of your choice. I chose to make Level 2 (MidSES) my reference category. When we choose a reference category in regression we set up a simple contrast.

Repeated Contrasts compare Level 1 to Level 2 and then Level 2 to Level 3 (then Level 3 to Level 4 and then level 4 to Level 5…).

Helmert Contrasts compare Level 1 to Levels 2 and 3 combined and then Level 2 to Level 3. (This comes in handy in the ILLCAUSE data set where we want to compare Healthy kids to Diabetic and Asthmatic kids and then compare Diabetic kids to Asthmatic kids.)

Unplanned Comparisons: Post Hoc Tests

ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.

Multiple Comparisons

Notice that there are only three t-tests going on here. Each test is reported twice for ease of reference.

Reading score
Bonferroni

(I) Socio Economic Status	(J) Socio Economic Status	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Low SES	Mid SES	-2.4664*	.68739	.001	-4.1135	-.8193
	High SES	-5.3065*	.63205	.000	-6.8211	-3.7920
Mid SES	Low SES	2.4664*	.68739	.001	.8193	4.1135
	High SES	-2.8402*	.48752	.000	-4.0084	-1.6720
High SES	Low SES	5.3065*	.63205	.000	3.7920	6.8211
	Mid SES	2.8402*	.48752	.000	1.6720	4.0084

Based on observed means.
The error term is Mean Square(Error) = 83.729.

*. The mean difference is significant at the 0.05 level.

In ANOVA, categorical variables are "factors" and the categories/values are "levels." Therefore the factor, SocioEconomicStatus, has three levels: LowSES (1), MidSES (2) and HighSES (3).

The more tests we conduct, the greater the chance of false positives (i.e., Type I Error). If we are conducting multiple tests we can adjust our alpha level. Here, we are conducting three different tests, so we will effectively divided our alpha level by three (alpha = 0.05/3 = 0.016), this is a Bonferroni correction, and it happens behind the scenes in our sampling distribution so we still say our alpha level is 0.05, but we note that we are making a Bonferroni adjustment.

Type I Error and Post Hoc Tests

ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.

Type I Error (False Positives): We err when we reject the null hypothesis and conclude that there is a relationship in the population when in fact there is no relationship in the population. This will happen (by design) for about 5% of our tests at alpha=.05 (when the null is true), unless we lower our alpha level or make adjustments (e.g., Bonferroni) for multiple comparisons.

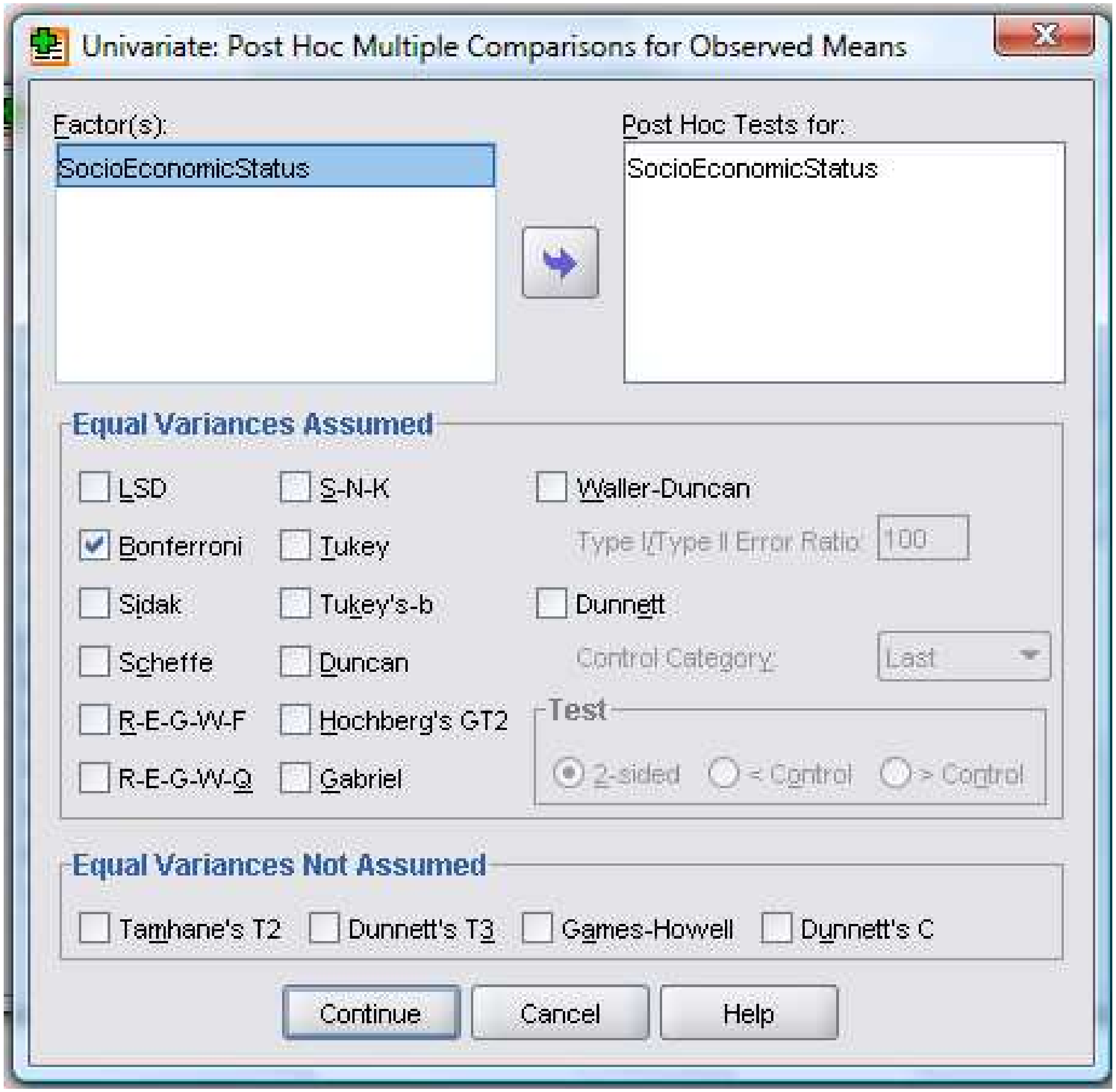
Type II Error (False Negatives): We err when we fail to reject the null hypothesis and withhold judgment about any relationship in the population when in fact there is a relationship in the population. Conventional wisdom says that Type I Error is four times worse than Type II Error, but I doubt that's true.

Planned Comparisons (Contrasts or A Priori Tests) are tests around which you designed your study. As soon as you get your results, you are going to make a bee line to your planned comparisons. You do not need to adjust your alpha level, because you are only looking at one (maybe two, maybe three…) tests for statistical significance. You can adjust for multiple comparisons if you plan to look at many tests. The key, however, is not how many tests the computer conducts but how many tests you conduct.

Unplanned Comparisons (Post Hoc Tests) are tests that come up along the way. When you have a polychotomous variable with five categories, e.g., RaceEthnicity: White, Black, Latino, Asian and Mixed, and you start comparing White students to Black students, and Asian students to Latino students, and White students to Latino students… then you are conducting 10 tests. Since confidence intervals succeed only 95% of the time, the chances that all 10 of your confidence intervals will succeed approaches a lowly 60% (95%*95%*95%*95%*95%*95%*95%*95%*95%*95%). This principal is why data-analytic fishing expeditions are so wrong. If you look for anything in your data, you will find something. Therefore, finding something is not interesting in and of itself. Interesting is when you look for something and you find it.

Choosing a Post Hoc Test

ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.



SPSS offers over a dozen adjustments for multiple comparison. They all do basically the same thing. I chose Bonferroni because it's the easiest to explain: divide your chosen alpha level by the number of comparisons. This is a very stringent (i.e., conservative) adjustment, almost certainly an eensy bit too conservative.

My only advice is: Never argue over which adjustment is better. If you are working with a devotee of a post hoc adjustment for multiple comparisons, go with their flow. Arguing about which post hoc test is better is like arguing about whether cats or dogs are better, whether soccer or football is better, or whether the deck chairs on the Titanic are better spaced equally or grouped together.

*Tukey's HSD, Cats, Football, Grouped

Plots

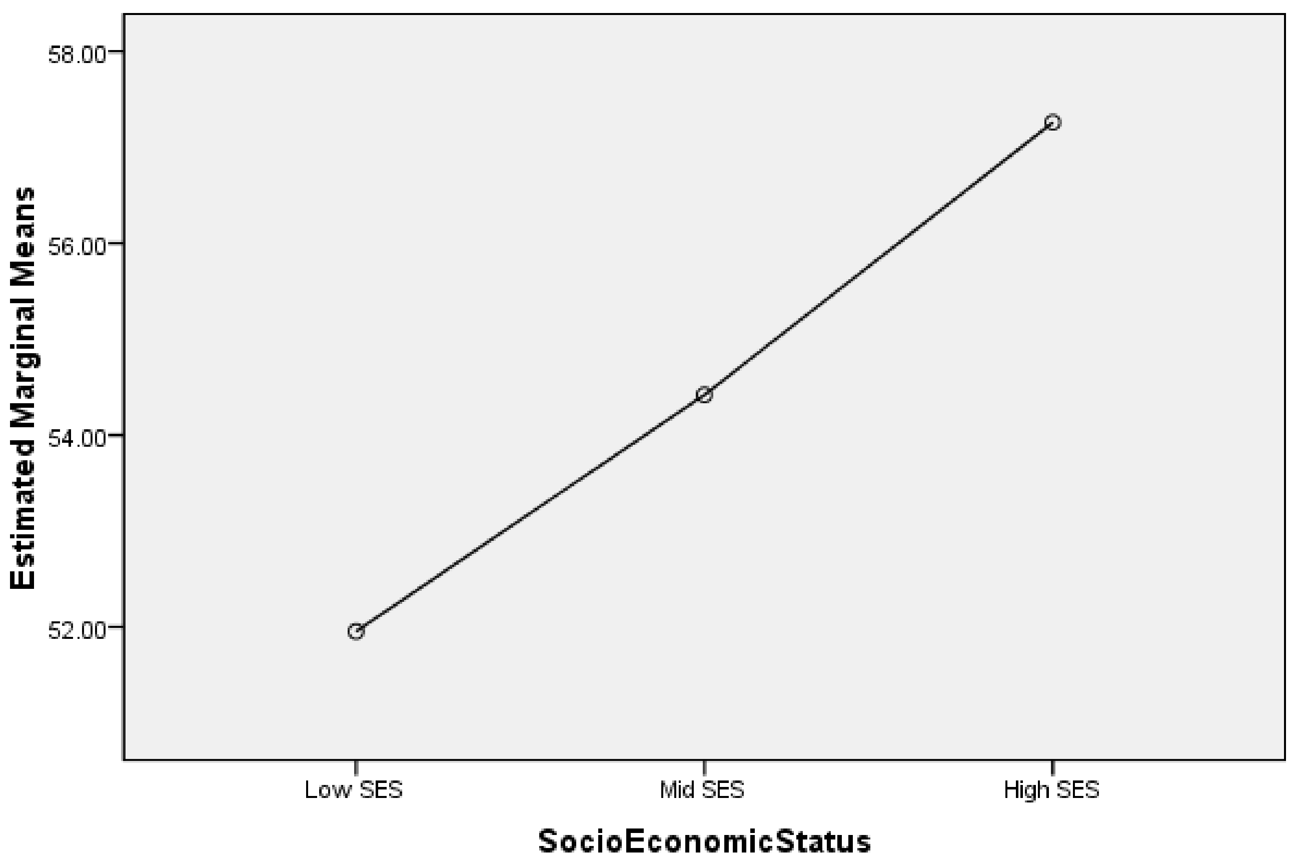
ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.

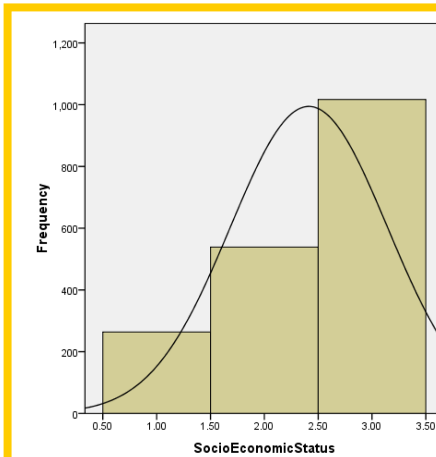
In Unit 10, we will talk about why the means are "marginal means."

Notice that the line is a little crooked. Our model allows it to be as crooked as it wants to be!

The skewness of a predictor often foreshadows linearity problems. This is an exception.

Estimated Marginal Means of Reading score



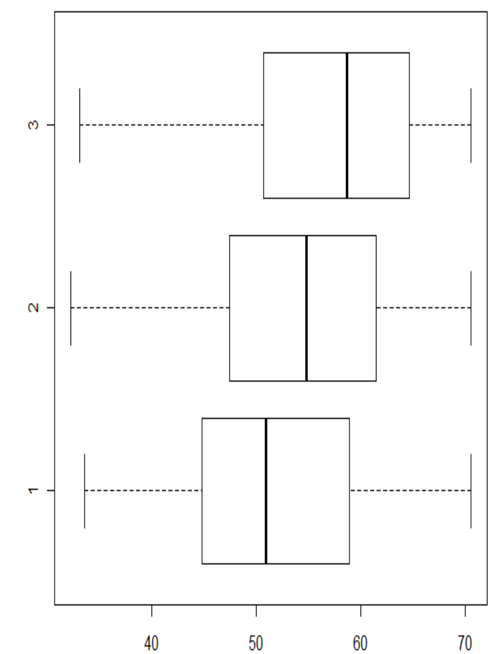


It might be helpful to note that our sample is restricted to four-year-college bound students, and such students tend to be high SES.

Contrasts, Post Hoc Tests and Plots (R)

ANOVA Perspective: Basic results partition the outcome variance. To dig deeper, use contrasts, post hoc tests and plots.

```
load("E:/Datasets/NELSBoys/nelsboys.rda")
attach(nelsboys)
# here is an R shortcut for polychotomies
# treat socioeconomicstatus as a factor, calling the factor whatever
ses.factor <- as.factor(socioeconomicstatus)
model.3 <- lm(read ~ ses.factor)
summary(model.3)
# when you include a factor in a model, R by default uses a dummy contrast
contrast(ses.factor) # will show the default dummy coding
# but you can change (i.e., re-assign) from the default contrast to Helmert
contrast(ses.factor) <- contr.helmert(3) # where the 3 indicates 3 levels
# if you want to switch back, you can change back
# R calls the default dummy coding “treatment coding”
# and you can specify the reference category by changing the base
contr.treatment(1, base = 1))
# for more see:,http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm
# for Post Hocs, get a matrix of Bonferroni adjusted p-values
pairwise.t.test(read, socioeconomicstatus, p.adj="bonferroni")
# here is more info: http://www.stat.wisc.edu/~yandell/st571/R/append12.pdf
# let’s create a boxplot for a visual
boxplot(read ~ socioeconomicstatus, horizontal=TRUE)
detach(nelsboys)
```




```
Call:
lm(formula = read ~ ses.factor)

Residuals:
    Min      1Q  Median      3Q     Max
-24.118  -6.723   0.762   7.202  18.598

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.5427     0.2483 219.674  < 2e-16 ***
ses.factor1   1.2332     0.3437   3.588 0.000342 ***
ses.factor2   1.3578     0.1492   9.098  < 2e-16 ***
```

```
Pairwise comparisons using t tests with pooled SD
data:  read and socioeconomicstatus
  1       2
2 0.0010  -
3 2.8e-16 2.0e-08
P value adjustment method: bonferroni
```

← With this Helmert contrast, we contrast the second level with the first (1.2332), and we contrast the third level with the average of the first two levels (1.3578). That’s what the p-values mean.

Helmert:

	x1	x2
lev.1	-1	-1
lev.2	1	-1
lev.3	0	2

To understand what the coefficients mean, you have to plug in the right numbers. E.g., for the mean of level 1, plug in -1 and -1. ↑

Interpreting Your Results

- Always start by searching HI-N-LO for assumption violations. At core, we are doing linear regression (i.e., applying the general linear model) whether we are doing t-tests or ANOVA, so the regression assumptions are relevant. Honestly, most people jump to the significance tests and then check their assumptions (I know I do), but you should feel appropriately guilty for not doing first things first, guilty enough to check your assumptions soon thereafter.
- Regression
  - Note the p-value (i.e., significance level) associated with the omnibus F-test to see if anything is going on in your model.
  - Note the p-values (i.e., significance levels) associated with your slope estimates to see where the action is in your model.
  - Interpret your statistically significant slopes.

- ANOVA
  - Note the p-value (i.e., significance level) associated with the omnibus F-test to see if anything is going on in your model.
  - In Two-Way ANOVA, which we discuss in Unit 10, the omnibus F-test is broken down into subtests, and these will give you a clue to where the action is.
  - Check your contrasts, post hoc comparisons and/or plots to see where the action is.

You have everything you need for Posthole 9. Practice is in back.

Dig the Post Hole (SPSS)

Unit 9 Post Hole:

Interpret the parameter estimates and F-test from regressing a continuous variable on a set of dummy variables.

Evidentiary material: regression output. In Los Angeles (circa 1980), interviewers from the Institute for Social Science Research at UCLA surveyed a multiethnic sample of 256 community members for an epidemiological study of depression (Afifi and Clark 1984). Reference category : NON-RELIGIOUS.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.230[a]	.053	.038	.79637

a. Predictors: (Constant), OTHER, JEWISH, CATHOLIC, PROTESTANT

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.924	4	2.231	3.518	.008[a]
	Residual	159.186	251	.634		
	Total	168.109	255			

a. Predictors: (Constant), OTHER, JEWISH, CATHOLIC, PROTESTANT

b. Dependent Variable: DEPRESS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.750	.110		6.791	.000	.532	.968
	PROTESTANT	-.306	.130	-.189	-2.352	.019	-.563	-.050
	CATHOLIC	-.163	.161	-.077	-1.011	.313	-.481	.154
	JEWISH	.207	.199	.073	1.036	.301	-.186	.599
	OTHER	.750	.574	.081	1.307	.192	-.380	1.880

a. Dependent Variable: DEPRESS

Here is the answer blank:

Yakkity yak yak yak, $F(df_{between}, df_{within})$ = xx.x, p = .xxx, η2 =.xxx.

Here is my answer:

There is a statistically significant relationship between depression and religion $F(4, 251) = 3.52$, $p = .008$, $\eta^2 = .053$.

Protestants tend to be less depressed than their non-religious counterparts ($p = .019$). There are no statistically significant differences in depression between non-religious subjects and subjects who self-identify as Catholic, Jewish, or other.

Do not trust these results until you've checked the regression assumptions! (They are ugly.) In general, don't trust any results until you've checked the assumptions or you trust that the researcher checked the assumptions. How do you know the researcher checked the assumptions? Look for clues.

Major clue that a researcher has checked the model assumptions: The researcher mentions and addresses one assumption concern somewhere. If a researcher has checked one assumption, she has probably checked most or all of the other assumptions.

Dig the Post Hole (R)

Unit 9 Post Hole:

Interpret the parameter estimates and F-test from regressing a continuous variable on a set of dummy variables.

Evidentiary material: regression output. In Los Angeles (circa 1980), interviewers from the Institute for Social Science Research at UCLA surveyed a multiethnic sample of 256 community members for an epidemiological study of depression (Afifi and Clark 1984). Reference category : NON-RELIGIOUS.

Here is the answer blank:

Yakkity yak yak yak, $F(df_{between}, df_{within})$ = xx.x, p = .xxx, η2 =.xxx.

Here is my answer:

There is a statistically significant relationship between depression and religion $F(4, 251) = 3.52$, $p = .008$, $\eta^2 = .053$.

Protestants tend to be less depressed than their non-religious counterparts ($p = .019$). There are no statistically significant differences in depression between non-religious subjects and subjects who self-identify as Catholic, Jewish, or other.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7500     0.1104   6.791    8e-11 ***
protestant   -0.3064     0.1302  -2.352   0.0194 *
catholic     -0.1630     0.1612  -1.011   0.3128
jewish        0.2065     0.1994   1.036   0.3014
other         0.7500     0.5738   1.307   0.1924
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.7964 on 251 degrees of freedom
Multiple R-squared: 0.05308, Adjusted R-squared: 0.03799
F-statistic: 3.518 on 4 and 251 DF,  p-value: 0.008157
```

Grab the F statistic here.

Regression with dummies is particularly easy in R, because if you identify a polychotomy as a factor, you can just include the factor in the model, and R will make the dummies for you. R will automatically use the lowest factor (i.e., smallest numerically or earliest alphabetically) as the reference category.

One-Way ANOVA (Live)

Simulating ANOVA tables

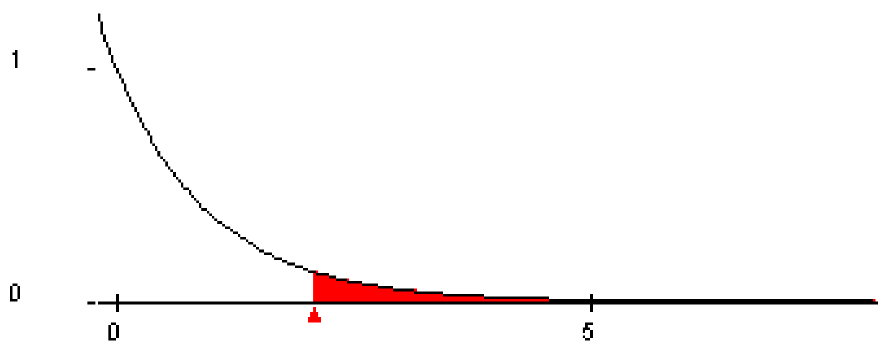
http://www.rossmanchance.com/applets/Anova/Anova.html

μ_1 21.0 — n_1 100

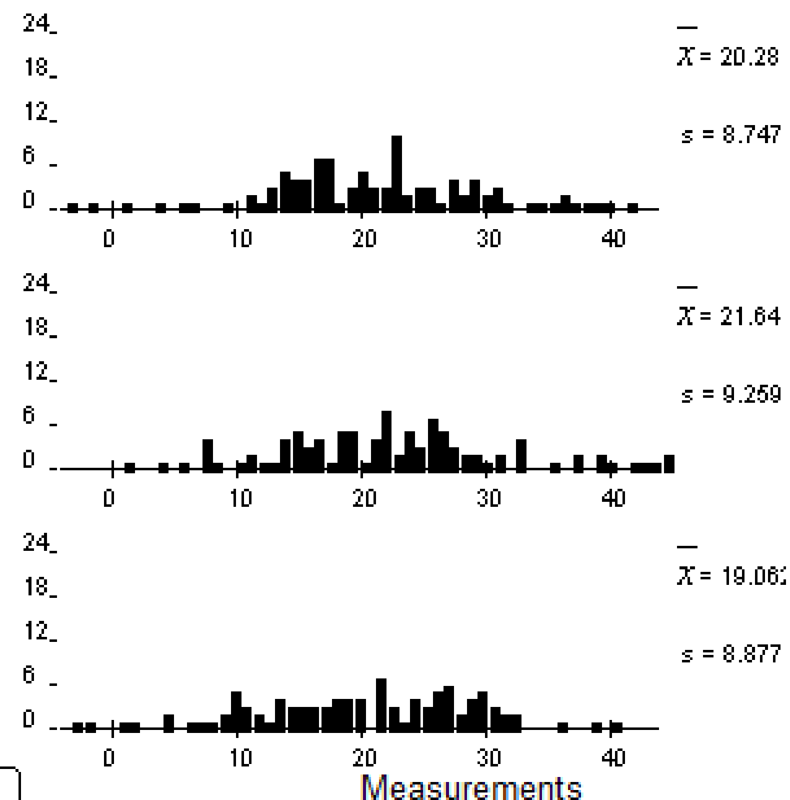
μ_2 22.0 — n_2 100

μ_3 18.0 — n_3 100

σ 9.0 — Draw Samples



Source	DF	SumSqr	MeanSqr	F	P-val
Groups	2	332.64	166.32	2.07	0.128
Error	297	23,863.21580.348			
Total	299	24,195.855			



DotPlot

BoxPlot

Reset

http://wise.cgu.edu/applets/Correl/correl.html
http://www.csustan.edu/ppa/llg/stat_demos.htm

Answering our Roadmap Question (Regression Perspective)

Unit 9: In the population, is there a relationship between reading and race?

$$Reading = \beta_0 + \beta_1 Asian + \beta_2 Latino + \beta_3 Black + \varepsilon$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.221[a]	.049	.049	8.35882

a. Predictors: (Constant), BLACK, ASIAN, LATINO

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	28016.721	3	9338.907	133.662	.000[a]
	Residual	544705.143	7796	69.870		
	Total	572721.864	7799			

a. Predictors: (Constant), BLACK, ASIAN, LATINO

b. Dependent Variable: READING

In our nationally representative sample of 7,800 8th graders, there is a statistically significant relationship between reading and race/ethnicity, F (3, 7796) = 133.7, p < .001. Based on 95% confidence intervals, the Black/White achievement gap is between 5.6 to 4.2 points. The Latino/White gap is between 5.0 and 3.8 points. The Asian/White gap favors Asians, and it is between 0.3 and 1.8 points.

We are using three confidence intervals. Should we use a Bonferroni adjustment? 95%*95%*95% = 86% which tells us that, over the course of our lives, in 14% of our triple confidence interval constructions, at least one of our three confidence intervals will miss the true population gap.

Coefficients[a]

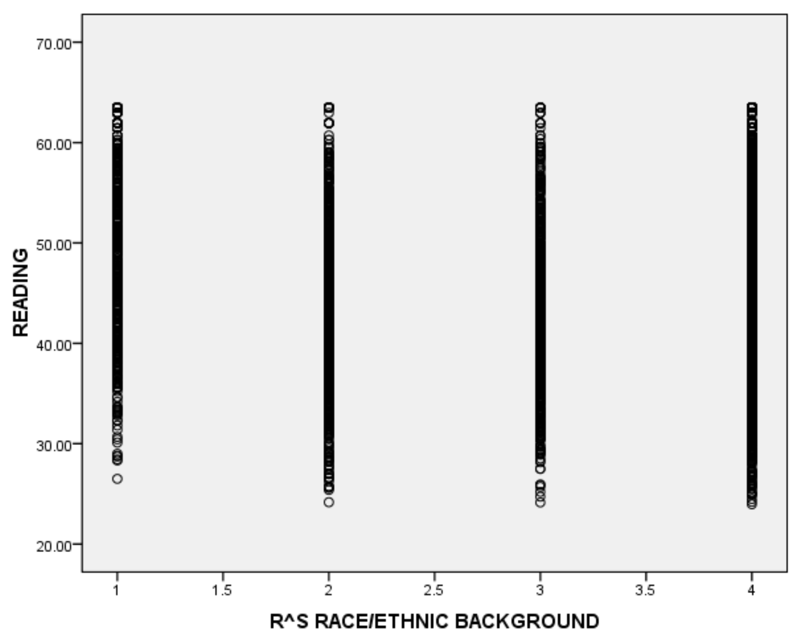
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.338	.110		438.242	.000	48.122	48.554
	ASIAN	1.034	.383	.030	2.697	.007	.283	1.786
	LATINO	-4.418	.306	-.161	-14.447	.000	-5.017	-3.818
	BLACK	-4.889	.339	-.161	-14.423	.000	-5.554	-4.225

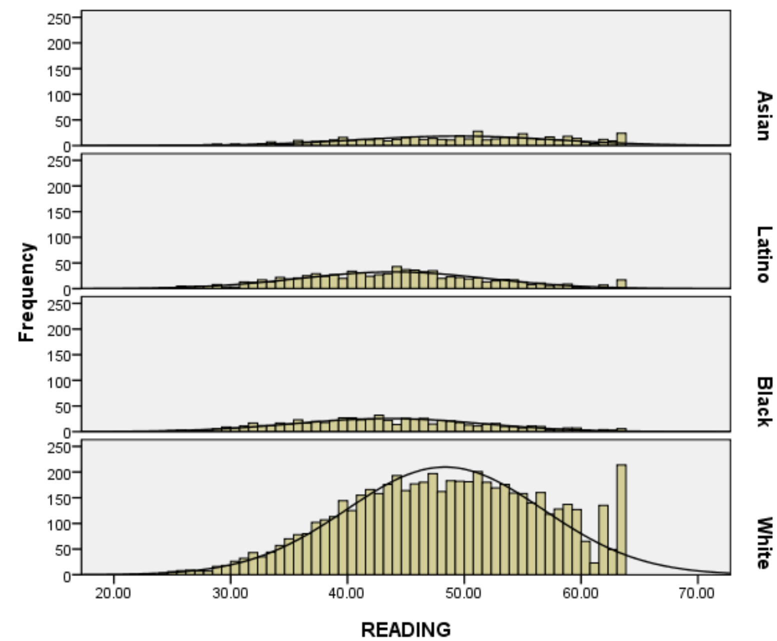
a. Dependent Variable: READING

Scatterplots vs. Histograms

For categorical predictors, you may find histograms more helpful than scatterplots. Note that you can "see" the histogram in the scatterplot if you rotate the scatterplot 90 degrees clockwise or you rotate the histograms 90 degrees counterclockwise.

All the regression assumptions from Unit 8 (HI-N-LO) apply to regression <u>and ANOVA</u> with polychotomies! (As with regression on dichotomies, the linearity assumption is a given, because we are only comparing two means at a time (where our reference category provides the common basis of comparison), and a straight line always passes through two points perfectly.





Answering our Roadmap Question (ANOVA Perspective—Basic)

Unit 9: In the population, is there a relationship between reading and race?

Between-Subjects Factors

		Value Label	N
R^S RACE/ETHNIC BACKGROUND	1	Asian	518
	2	Latino	859
	3	Black	680
	4	White	5743

In our nationally representative sample of 7,800 8th graders, there is a statistically significant relationship between reading and race/ethnicity, $F(3, 7796) = 133.7$, $p < .001$. In our sample, race/ethnicity predicts 5% of the variation in reading scores.

Tests of Between-Subjects Effects

Dependent Variable:READING

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	28016.721[a]	3	9338.907	133.662	.000	.049
Intercept	7227666.576	1	7227666.576	103444.752	.000	.930
RACE	28016.721	3	9338.907	133.662	.000	.049
Error	544705.143	7796	69.870			
Total	1.817E7	7800				
Corrected Total	572721.864	7799				

a. R Squared = .049 (Adjusted R Squared = .049)

Answering our Roadmap Question (ANOVA Perspective—Digging Deeper)

Unit 9: In the population, is there a relationship between reading and race?

Post Hoc Tests

Multiple Comparisons

READING
Bonferroni

(I) R^S RACE /ETHNIC...	(J) R^S RACE /ETHNIC...	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Asian	Latino	5.4520*	.46500	.000	4.2249	6.6790
	Black	5.9236*	.48748	.000	4.6372	7.2100
	White	1.0343*	.38347	.042	.0223	2.0462
Latino	Asian	-5.4520*	.46500	.000	-6.6790	-4.2249
	Black	.4716	.42906	1.000	-.6606	1.6039
	White	-4.4177*	.30579	.000	-5.2246	-3.6107
Black	Asian	-5.9236*	.48748	.000	-7.2100	-4.6372
	Latino	-.4716	.42906	1.000	-1.6039	.6606
	White	-4.8893*	.33899	.000	-5.7839	-3.9947
White	Asian	-1.0343*	.38347	.042	-2.0462	-.0223
	Latino	4.4177*	.30579	.000	3.6107	5.2246
	Black	4.8893*	.33899	.000	3.9947	5.7839

Based on observed means.
The error term is Mean Square(Error) = 69.870.

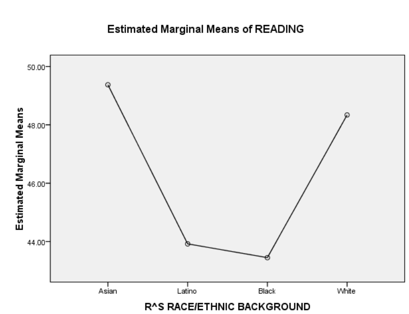
*. The mean difference is significant at the .05 level.

Contrast Results (K Matrix)

R^S RACE/ETHNIC BACKGROUND Simple Contrast[a]			Dependent Variable READING
Level 1 vs. Level 4	Contrast Estimate		1.034
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		1.034
	Std. Error		.383
	Sig.		.007
	95% Confidence Interval for Difference	Lower Bound	.283
		Upper Bound	1.786
Level 2 vs. Level 4	Contrast Estimate		-4.418
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-4.418
	Std. Error		.306
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-5.017
		Upper Bound	-3.818
Level 3 vs. Level 4	Contrast Estimate		-4.889
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-4.889
	Std. Error		.339
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-5.554
		Upper Bound	-4.225

a. Reference category = 4

Profile Plots



Contrasts: Notice that the "Simple Contrast" gives us the same information as our regression coefficients, standard errors and confidence intervals.

Post Hoc Tests: We are making 6 comparisons, so a Bonferroni adjustment is probably in order. All the pairwise differences are statistically significant (with a Bonferroni adjusted alpha level of .008 (.05/6)), except for the difference between Latino students and Black students.

Plots: Notice that the lines are silly. They seem to imply an order among ASIAN, LATINO, BLACK and WHITE, but RACE is not an ordinal variable. Rather, RACE is a nominal variable, so look beyond the lines.

Answering our Roadmap Question (Regression Perspective)

Unit 9: In the population, is there a relationship between reading and race?

$$Reading = \beta_0 + \beta_1 Asian + \beta_2 Latino + \beta_3 Black + \varepsilon$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.221[a]	.049	.049	8.35882

a. Predictors: (Constant), BLACK, ASIAN, LATINO

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	28016.721	3	9338.907	133.662	.000[a]
	Residual	544705.143	7796	69.870		
	Total	572721.864	7799			

a. Predictors: (Constant), BLACK, ASIAN, LATINO

b. Dependent Variable: READING

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.338	.110		438.242	.000	48.122	48.554
	ASIAN	1.034	.383	.030	2.697	.007	.283	1.786
	LATINO	-4.418	.306	-.161	-14.447	.000	-5.017	-3.818
	BLACK	-4.889	.339	-.161	-14.423	.000	-5.554	-4.225

a. Dependent Variable: READING

In our nationally representative sample of 7,800 8th graders, there is a statistically significant relationship between reading and race/ethnicity achievement, F (3, 7796) = 133.7, p < .001. Based on 95% confidence intervals, the Black/White achievement gap is between 5.6 to 4.2 points. The Latino/White gap is between 5.0 and 3.8 points. The Asian/White gap favors Asians, and it is between 0.3 and 1.8 points.

We are using three confidence intervals. Should we use a Bonferroni adjustment? 95%*95%*95% = 86% which tells us that, over the course of our lives, in 14% of our triple confidence interval constructions, at least one of our three confidence intervals will miss the true population gap.

Unit 9 Appendix: Key Concepts

Linear regression is a flexible tool that subsumes t-tests and ANOVA. All three data analytic tool fall under the rubric of the general linear model.

Nevertheless, t-tests come in useful flavors that can be difficult to replicate using regression.

- **Special t-tests can correct standard errors for inequality of variances, heteroscedasticity.**
- **Special t-tests can adjust for non-independent observations, for example, repeated measures.**

ANOVA has useful output that to replicate in regression can require specialized programming skills.

- **Contrasts**
- **Post Hoc Comparisons**
- **Plots**

Unit 9 Appendix: Key Interpretations

There is a statistically significant relationship between socioeconomic status and reading scores, $F(2,1817) = 42.155$, $p < .001$. The null hypothesis is that there is no relationship in the population. We reject the null hypothesis based on a p-value of less than .05. We conclude that there is a relationship in the population.

The difference that we observe in our sample, five points, is statistically significant ($p < 0.001$). We estimate that the Latino/Anglo reading gap is between 6.5 and 3.5 points in the population of four-year-college bound boys. We emphasize that we are predicting group averages, not individuals. The best Latino reader in our sample reads as well as the best Anglo reader, and the worst Latino reader in our sample reads better than the worst Anglo reader.

In our sample of 1820 four-year-college bound boys, we observe a statistically significant relationship between reading and socioeconomic status, $F(2,1817) = 42.2$, $p < 0.001$. All pairwise comparisons are statistically significant based on a Bonferroni adjusted alpha level of 0.017 (0.05/3). Boys of high SES ($M = 57.2$) tended to read better than boys of middle SES ($M = 54.4$) who tended to read better than boys of low SES ($M = 51.2$). Nevertheless, we note that there is much variation within groups as evidenced by our R2 statistic of 0.044, which tells us that 95.6% percent of the variation in reading achievement remains unpredicted by socioeconomic status.

In our nationally representative sample of 7,800 8th graders, there is a statistically significant relationship between reading achievement and race/ethnicity, $F(3, 7796) = 133.7$, $p < .001$. Based on 95% confidence intervals, the Black/White achievement gap is between 5.6 to 4.2 points. The Latino/White gap is between 5.0 and 3.8 points. The Asian/White gap favors Asians, and it is between 0.3 and 1.8 points. In our sample, race/ethnicity predicts 5% of the variation in reading scores.

Unit 9 Appendix: Key Terminology

- The key to addressing any null hypothesis is to consider the sampling distribution for your statistic. If you took a thousand (equally sized) random samples from the population, and you calculated your statistic for each sample, how would the statistics distribute themselves? Whereas means and slopes form a normal distribution (or t distribution), F statistics form a positively skewed distribution the exact shape of which depends on not only the degrees of freedom of the subjects but also the degrees of freedom of the variables. Once you have your sampling distribution, you can set it at zero (as per the null hypothesis) and observe if your statistic is far enough away from zero (based on your alpha level) to reject the null hypothesis.
- **Type I Error (False Positives):** See Slide 34.
- **Type II Error (False Negatives):** See Slide 34.
- **Planned Comparisons (Contrasts or A Priori Tests):** See Slide 34.
- **Unplanned Comparisons (Post Hoc Tests):** See Slide 34.
- **In ANOVA, categorical variables are "Factors" and the categories/values are "Levels." Therefore the factor, SocioEconomicStatus, has three levels: LowSES (1), MidSES (2) and HighSES (3).**
- **A set of Dummy Variables is a set of dichotomous predictors coded 0/1 that represent a polychotomous predictor for the sake of linear regression.**
- **The omnibus F-test tests the null hypothesis that there is no relationship in the population between any of the predictors individually or combined. In other words, it tests the null hypothesis that in the population, the R^2 statistic is zero.**

Unit 5 Appendix: Math (reprised)

Every individual observation gets three squares:

The blue square represents the squared difference between the observed outcome for the individual and the mean of the outcome.

$$(Y_i - \bar{Y})^2$$

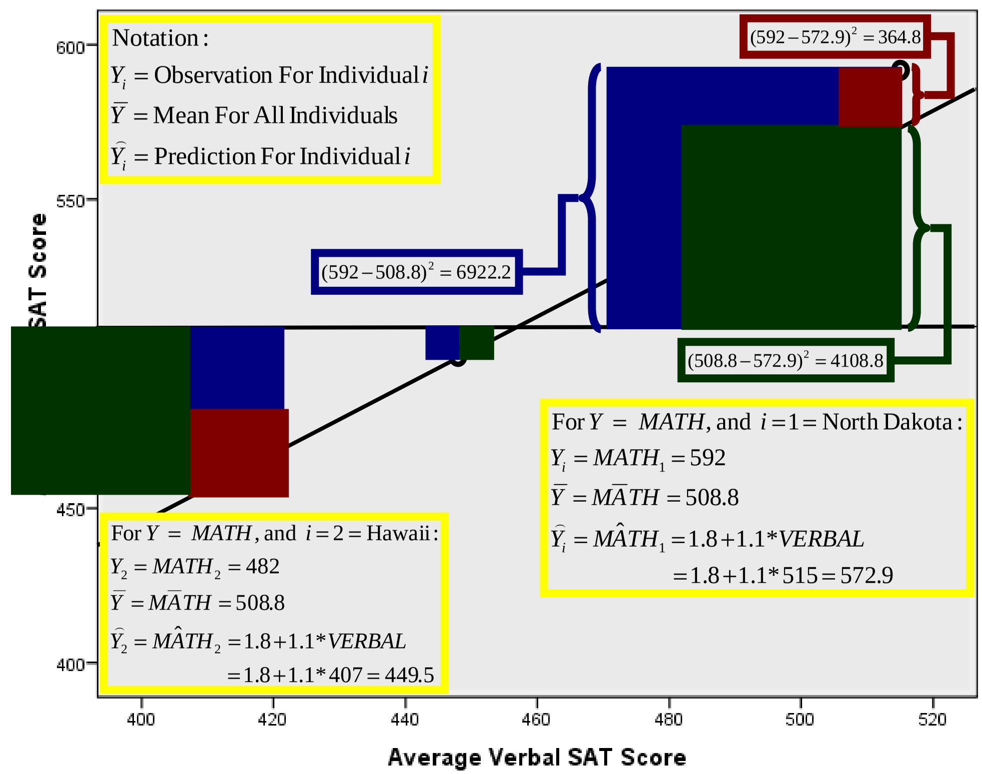
The red square represents the squared difference between the observed outcome for the individual and the predicted outcome for the individual.

$$(Y_i - \hat{Y}_i)^2$$

The green square represents the squared difference between the mean of the outcome and the predicted outcome for the individual.

$$(\bar{Y} - \hat{Y}_i)^2$$

Cool Algebraic Fact: Because of the way we fit our regression line, all the blue squares combined equal all the red squares combined plus all the green squares combined.
http://en.wikipedia.org/wiki/Sum_of_squares



$$\text{Sum of Squares Total} = SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$\text{Sum of Squares Residual/Error} = SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\text{Sum of Squares Regression/Model} = SSM = \sum_{i=1}^{n} (\bar{Y} - \hat{Y}_i)^2$$

Cool Algebraic Fact: SST = SSE + SSM

Don't be afraid of capital sigma (Σ). It is the capital Greek letter S, and it stands for "sum." It just means "add 'em up"! You calculate SST for Post Hole 3.

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST}$$

The R^2 statistic is a goodness of fit statistic:

$$R^2 = 1 - \frac{BAD}{BASELINE} = \frac{GOOD}{BASELINE}$$

Unit 9 Appendix: Math

Intuitive Representations of the F statistic:

$$F = \frac{\text{Predicted Variation}}{\text{Unpredicted Variation}} = \frac{\text{Good Mean Square}}{\text{Bad Mean Square}} = \frac{\text{Signal}}{\text{Noise}} = \text{Want Big} = \frac{\text{Want Big}}{\text{Want Small}}$$

Technical But Still Verbal Representations of the F statistic:

$$F = \frac{\text{Regression Mean Square}}{\text{Residual Mean Square}} = \frac{\text{Model Mean Square}}{\text{Error Mean Square}} = \frac{\text{Between-Groups Mean Square}}{\text{Within-Groups Mean Square}}$$

Formal Representation of the F statistic:

$$F = \frac{\dfrac{\sum_{i=1}^{n}(\bar{Y} - \hat{Y}_i)^2}{k-1}}{\dfrac{\sum_{i=1}^{n}(Y_i - Y_i)^2}{n-k}}$$

Notation:

n = Number of Observations

k = Number of Parameters/Groups

Y_i = Observation For Individual i

$\bar{Y}$ = Mean For All Individuals

$\hat{Y}_i$ = Prediction For Individual i

What is a "Mean Square"?

For Post Hole 3, we calculate a mean square, and we call it "variance." In general, to calculate a mean, we add up a number of things and divide by the number of things. To calculate a mean square, we add up (i.e., sum) a number of squares and divide by the degrees of freedom.

The numerator of the F statistic is the regression sum of squares divided by the degrees of freedom.

The denominator of the F statistic is the residual sum of squares divided by the degrees of freedom.

Once you have your F statistic, you compare it to the "critical value" from an F-distribution with the appropriate degrees of freedom (k-1, n-k). A critical value is the cut-off based on your alpha level. If your observed F statistic is greater than the critical value, you reject the null hypothesis. On the next slide is a F table with critical values.

Unit 9 Appendix: Math

Critical F Values for Alpha = .05

Denominator Degrees of Freedom (n-k)	Numerator Degrees of Freedom (k-1)									
	1	2	3	4	5	6	7	8	9	10
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.17
40	4.09	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.98	1.93

Unit 9 Appendix: Math

Question: We leave out one of our indicators so that we do not include redundant information, but what's wrong with including redundant information?

Answer: If we conclude redundant information, there is no longer a unique fitted model, and the computer can't choose between all the equally good fitted models, so it blows up.

Given a dataset and estimation method, there is only one fitted model for this theoretical model:

$$Reading = \beta_0 + \beta_1 Asian + \beta_2 Latino + \beta_3 Black + \varepsilon$$

$$\hat{Reading} = 48 + 1 * Asian - 4 * Latino - 5 * Black$$

There are many equally good fitted models for this theoretical model:

$$Reading = \beta_0 + \beta_1 Asian + \beta_2 Latino + \beta_3 Black + \beta_4 White + \varepsilon$$

$$\hat{Reading} = 0 + 49 * Asian + 44 * Latino + 43 * Black + 48 * White$$

$$\hat{Reading} = 1 + 48 * Asian + 43 * Latino + 42 * Black + 47 * White$$

$$\hat{Reading} = 40 + 9 * Asian + 4 * Latino + 3 * Black + 8 * White$$

So, to specify a theoretical model with a unique fitted model, we need to drop something. We drop one of the dummies/indicators. We could alternatively drop the y-intercept, effectively forcing the y-intercept to be zero, which would reduce the three above possibilities to only one possibility, the first.

Unit 9 Appendix: SPSS Syntax (Part I)

```
***********************************************************.
*Univariate Exploration.
***********************************************************.
GRAPH
  /HISTOGRAM(NORMAL)=Read.
GRAPH
  /HISTOGRAM(NORMAL)=Latino.
GRAPH
  /HISTOGRAM(NORMAL)=SocioEconomicStatus.
FREQUENCIES VARIABLES=Read Latino
  /FORMAT=NOTABLE
  /NTILES=4
  /STATISTICS=STDDEV MEAN
  /ORDER=ANALYSIS.
FREQUENCIES VARIABLES=SocioEconomicStatus Latino.
```

Unit 9 Appendix: SPSS Syntax (Part II)

```
**********************************************************.
*Dichotomous Predictor.
**********************************************************.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER Latino.

GRAPH
  /SCATTERPLOT(BIVAR)=Latino WITH Read
  /MISSING=LISTWISE.

* This gets us a t-test, where we compare the two groups of Latino: 0 and 1.
T-TEST GROUPS= Latino(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=Read
  /CRITERIA=CI(.9500).

*This gives us a one-way ANOVA.
UNIANOVA Read BY Latino
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=ETASQ
  /CRITERIA=ALPHA(0.05)
  /DESIGN=Latino.
```

Unit 9 Appendix: SPSS Syntax (Part III)

```
**********************************************************.
*Polychotomous Predictor.
**********************************************************.

Compute LowSES  = 0.
If (SocioEconomicStatus = 1) LowSES = 1.
Compute HighSES  = 0.
If (SocioEconomicStatus = 3) HighSES = 1.
Execute.

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Read
  /METHOD=ENTER LowSES HighSES.

GRAPH
  /SCATTERPLOT(BIVAR)=SocioEconomicStatus WITH Read
  /MISSING=LISTWISE.

UNIANOVA Read BY SocioEconomicStatus           *This gives us a one-way ANOVA.
  /CONTRAST(SocioEconomicStatus)=Simple (2)    *This gives us a simple contrast with 2 as the reference category.
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=ETASQ
  /POSTHOC=SocioEconomicStatus(BONFERRONI)  *This gives us Bonferroni adjusted pairwise comparisons.
  /PLOT=PROFILE(SocioEconomicStatus)
  /CRITERIA=ALPHA(0.05)
  /DESIGN=SocioEconomicStatus.
```

Unit 9 Appendix: R Syntax

```
load("E:/Datasets/NELSBoys/nelsboys.rda")
attach(nelsboys)

hist(read)
summary(read)
summary(latino)
plot(read~latino)
boxplot(read~latino)
my.model <-  lm(read~latino)
summary(my.model)
anova(my.model)
t.test(read~latino, var.equal=FALSE)

boxplot(read ~ socioeconomicstatus, horizontal=TRUE)
# convert the polychotomy to dichotomies (dummies)
low.ses <- as.numeric(socioeconomicstatus==1)
mid.ses <- as.numeric(socioeconomicstatus==2)
high.ses <- as.numeric(socioeconomicstatus==3)
model.2 <- lm(read ~ low.ses + high.ses) # note that the reference category is level 2 since we left it out
summary(model.2)
anova(model.2)
# convert the polychotomy to a factor
ses.factor <- as.factor(socioeconomicstatus)
model.3 <- lm(read ~ ses.factor) # note that the reference category is level 1 by default
summary(model.3)
anova(model.3)

detach(nelsboys)
```

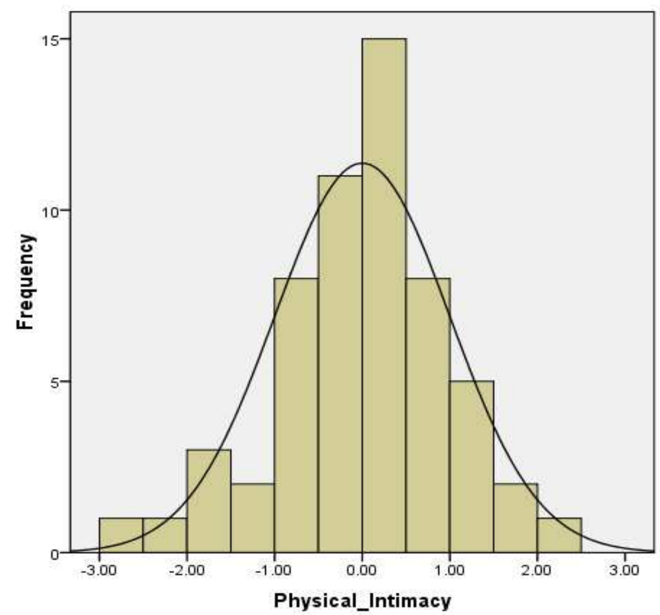
Perceived Intimacy of Adolescent Girls (Intimacy.sav)

- **Overview: Dataset contains self-ratings of the intimacy that adolescent girls perceive themselves as having with: (a) their mother and (b) their boyfriend.**
- **Source: HGSE thesis by Dr. Linda Kilner entitled Intimacy in Female Adolescent's Relationships with Parents and Friends (1991). Kilner collected the ratings using the Adolescent Intimacy Scale.**
- **Sample: 64 adolescent girls in the sophomore, junior and senior classes of a local suburban public school system.**
- **Note on Physical_Intimacy (with boyfriend): This is a composite variable based on a principle components analysis. Girls who score high on Physical_Intimacy scored high on (1) Physical Affection and (2) Mutual Caring, but low on (3) Risk Vulnerability and (4) Resolve Conflicts, regardless of (5) Trust and (6) Self Disclosure.**
- **Variables:**

(Physical_Intimacy)
Physical Intimacy With Boyfriend—see above

(RiskVulnerabilityWMom)
1=Tend to Risk Vulnerability with Mom, 0=Not

(ResolveConflictWMom)
1=Tend to Resolve Conflict with Mom, 0=Not



Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.071[a]	.005	-.013	1.00647

a. Predictors: (Constant), RiskVulnerabilityWMom

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.286	1	.286	.283	.597[a]
	Residual	55.714	55	1.013		
	Total	56.000	56			

a. Predictors: (Constant), RiskVulnerabilityWMom

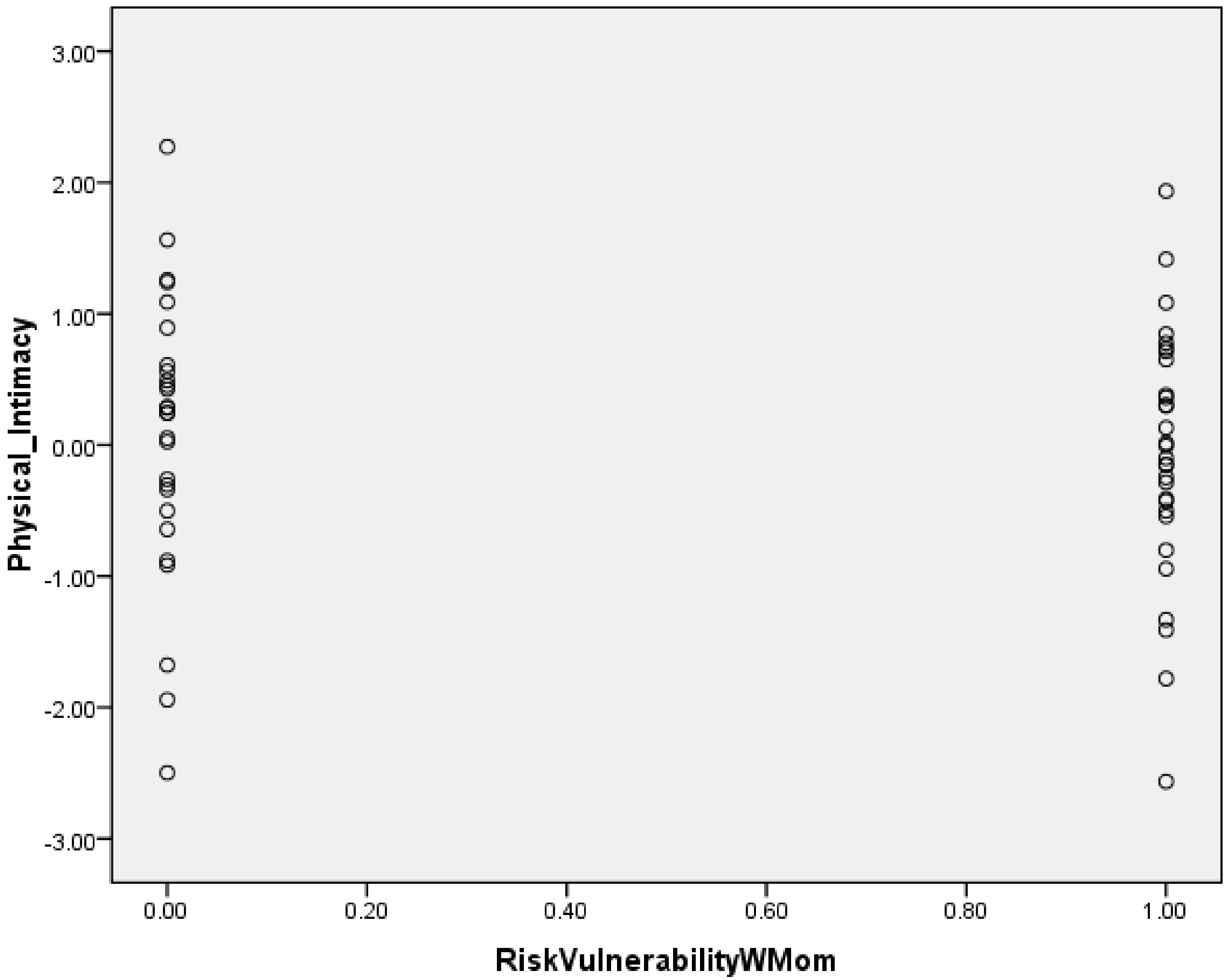
b. Dependent Variable: Physical_Intimacy

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.075	.194		.386	.701	-.313	.463
	RiskVulnerabilityWMom	-.142	.267	-.071	-.532	.597	-.677	.393

a. Dependent Variable: Physical_Intimacy

Perceived Intimacy of Adolescent Girls (Intimacy.sav)



Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Univariate Analysis of Variance

Between-Subjects Factors

		N
RiskVulnerabilityWMom	0	27
	1	30

Tests of Between-Subjects Effects

Dependent Variable:Physical_Intimacy

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.286[a]	1	.286	.283	.597
Intercept	.001	1	.001	.001	.978
RiskVulnerabilityWMom	.286	1	.286	.283	.597
Error	55.714	55	1.013		
Total	56.000	57			
Corrected Total	56.000	56			

a. R Squared = .005 (Adjusted R Squared = -.013)

Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.096[a]	.009	-.009	1.00437

a. Predictors: (Constant), ResConflictWMom

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.518	1	.518	.514	.477[a]
	Residual	55.482	55	1.009		
	Total	56.000	56			

a. Predictors: (Constant), ResConflictWMom

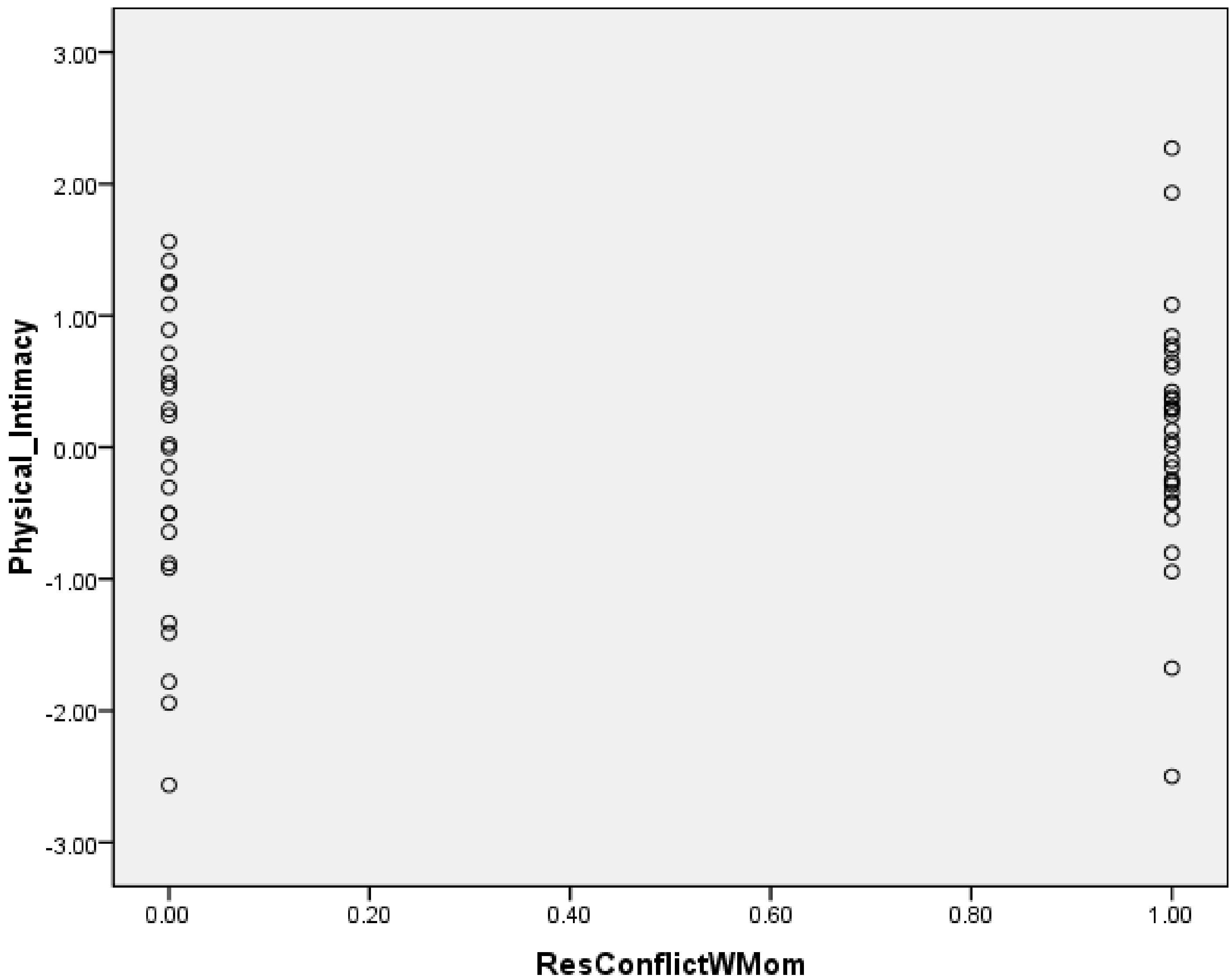
b. Dependent Variable: Physical_Intimacy

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.104	.197		-.529	.599	-.499	.291
	ResConflictWMom	.191	.267	.096	.717	.477	-.344	.727

a. Dependent Variable: Physical_Intimacy

Perceived Intimacy of Adolescent Girls (Intimacy.sav)



Perceived Intimacy of Adolescent Girls (Intimacy.sav)

Univariate Analysis of Variance

Between-Subjects Factors

		N
ResConflictWMom	0	26
	1	31

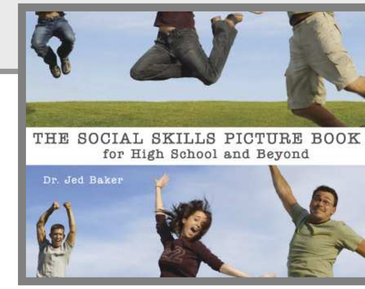
Tests of Between-Subjects Effects

Dependent Variable:Physical_Intimacy

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.518[a]	1	.518	.514	.477
Intercept	.004	1	.004	.004	.950
ResConflictWMom	.518	1	.518	.514	.477
Error	55.482	55	1.009		
Total	56.000	57			
Corrected Total	56.000	56			

a. R Squared = .009 (Adjusted R Squared = -.009)

High School and Beyond (HSB.sav)



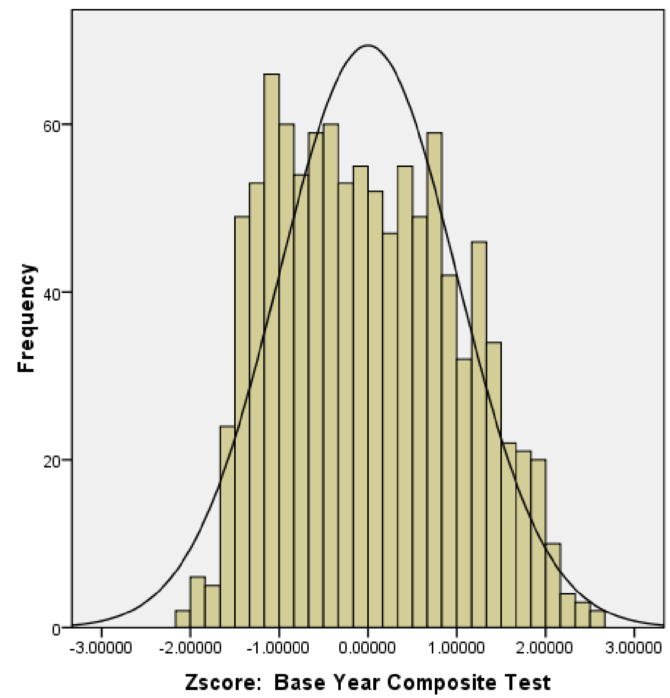
- **Overview: High School & Beyond – Subset of data focused on selected student and school characteristics as predictors of academic achievement.**
- **Source: Subset of data graciously provided by Valerie Lee, University of Michigan.**
- **Sample: This subsample has 1044 students in 205 schools. Missing data on the outcome test score and family SES were eliminated. In addition, schools with fewer than 3 students included in this subset of data were excluded.**
- **Variables:**

(ZBYTest)	Standardized Base Year Composite Test Score
(Sex)	1=Female, 0=Male
(RaceEthnicity)	Students Self-Identified Race/Ethnicity 1=White/Asian/Other, 2=Black, 3=Latino/a

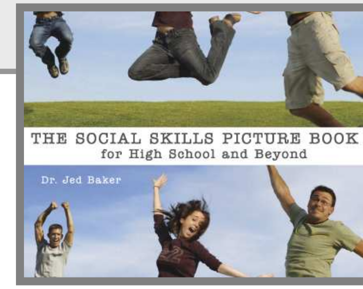
Dummy Variables for RaceEthnicity:

(Black)	1=Black, 0=Else
(Latin)	1=Latino/a, 0=Else

*Note that we will use RaceEthnicity=1, White/Asian/Other, as our reference category.



High School and Beyond (HSB.sav)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.158[a]	.025	.024	.98785042

a. Predictors: (Constant), 1 = Female, 0 = Other

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	26.166	1	26.166	26.814	.000[a]
	Residual	1016.834	1042	.976		
	Total	1043.000	1043			

a. Predictors: (Constant), 1 = Female, 0 = Other

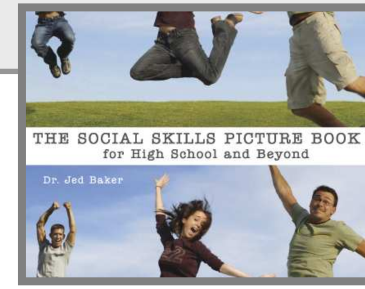
b. Dependent Variable: Zscore: Base Year Composite Test

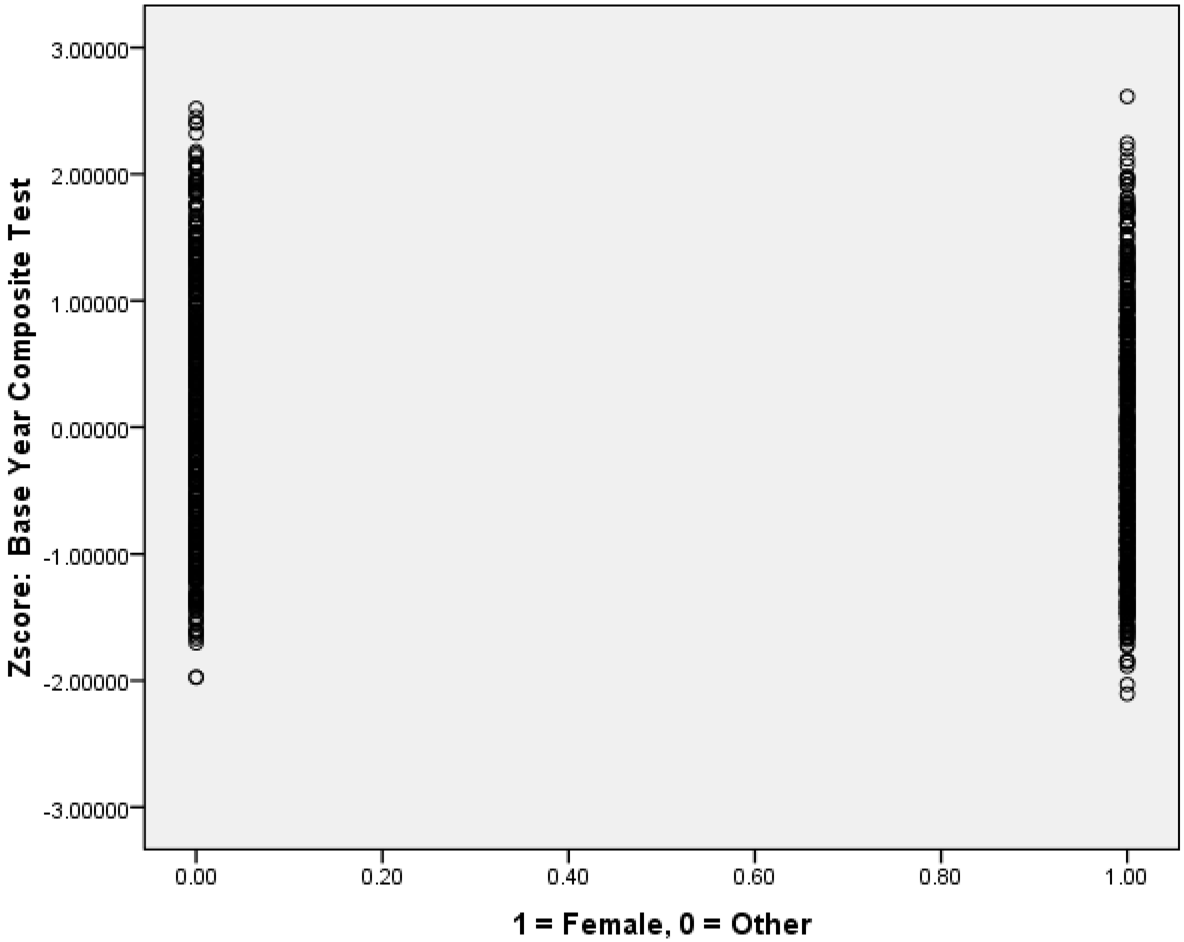
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.177	.046		3.856	.000	.087	.267
	1 = Female, 0 = Other	-.319	.062	-.158	-5.178	.000	-.439	-.198

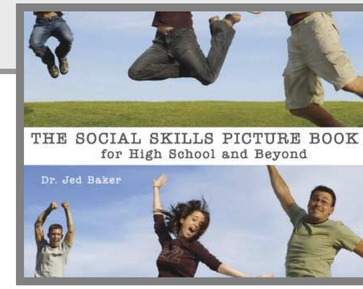
a. Dependent Variable: Zscore: Base Year Composite Test

High School and Beyond (HSB.sav)





High School and Beyond (HSB.sav)



Between-Subjects Factors

		Value Label	N
1 = Female, 0 = Other	0	Male	465
	1	Female	579

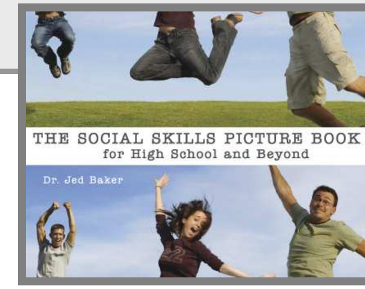
Tests of Between-Subjects Effects

Dependent Variable:Zscore: Base Year Composite Test

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	26.166[a]	1	26.166	26.814	.000
Intercept	.312	1	.312	.320	.572
Sex	26.166	1	26.166	26.814	.000
Error	1016.834	1042	.976		
Total	1043.000	1044			
Corrected Total	1043.000	1043			

a. R Squared = .025 (Adjusted R Squared = .024)

High School and Beyond (HSB.sav)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.432[a]	.187	.185	.90253787

a. Predictors: (Constant), 1 = Latino/a, 0 = Other, 1 = Black, 0 = Other

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	195.028	2	97.514	119.711	.000[a]
	Residual	847.972	1041	.815		
	Total	1043.000	1043			

a. Predictors: (Constant), 1 = Latino/a, 0 = Other, 1 = Black, 0 = Other

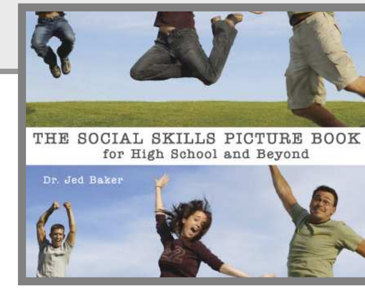
b. Dependent Variable: Zscore: Base Year Composite Test

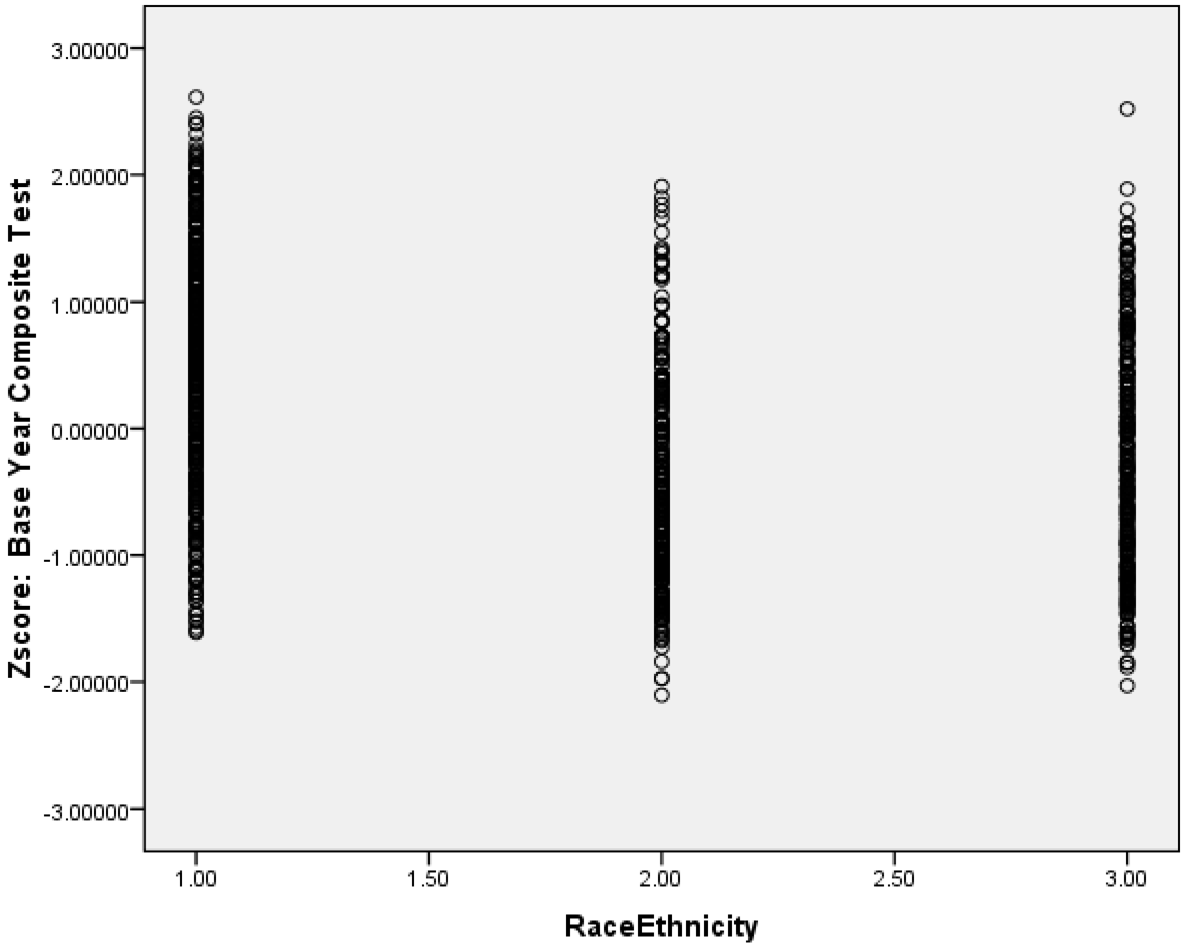
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.501	.043		11.565	.000	.416	.586
	1 = Black, 0 = Other	-.978	.068	-.442	-14.423	.000	-1.111	-.845
	1 = Latino/a, 0 = Other	-.741	.067	-.339	-11.056	.000	-.873	-.610

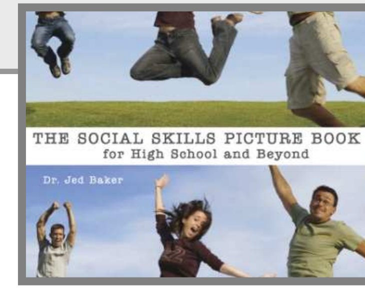
a. Dependent Variable: Zscore: Base Year Composite Test

High School and Beyond (HSB.sav)





High School and Beyond (HSB.sav)



Between-Subjects Factors

		Value Label	N
RaceEthnicity	1	White/Asian/Other	434
	2	Black	299
	3	Latino/a	311

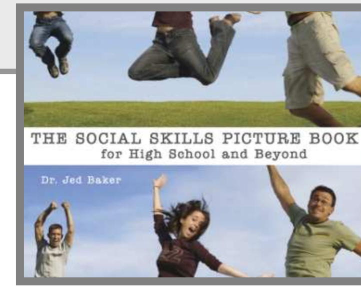
Tests of Between-Subjects Effects

Dependent Variable:Zscore: Base Year Composite Test

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	195.028[a]	2	97.514	119.711	.000
Intercept	5.292	1	5.292	6.496	.011
RaceEthnicity	195.028	2	97.514	119.711	.000
Error	847.972	1041	.815		
Total	1043.000	1044			
Corrected Total	1043.000	1043			

a. R Squared = .187 (Adjusted R Squared = .185)

High School and Beyond (HSB.sav)

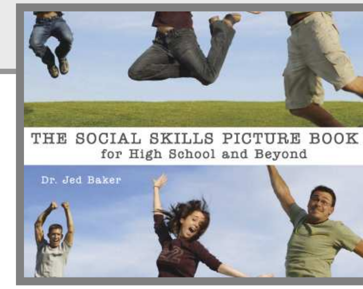


Contrast Results (K Matrix)

RaceEthnicity Simple Contrast[a]			Dependent ... Zscore: Base Year Composite Test
Level 2 vs. Level 1	Contrast Estimate		-.978
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.978
	Std. Error		.068
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-1.111
		Upper Bound	-.845
Level 3 vs. Level 1	Contrast Estimate		-.741
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.741
	Std. Error		.067
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-.873
		Upper Bound	-.610

a. Reference category = 1

High School and Beyond (HSB.sav)



Multiple Comparisons

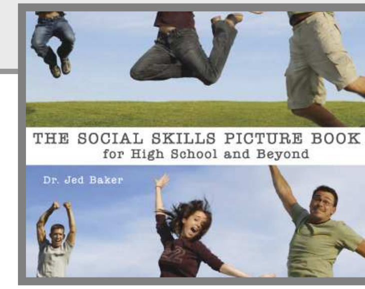
Zscore: Base Year Composite Test
Bonferroni

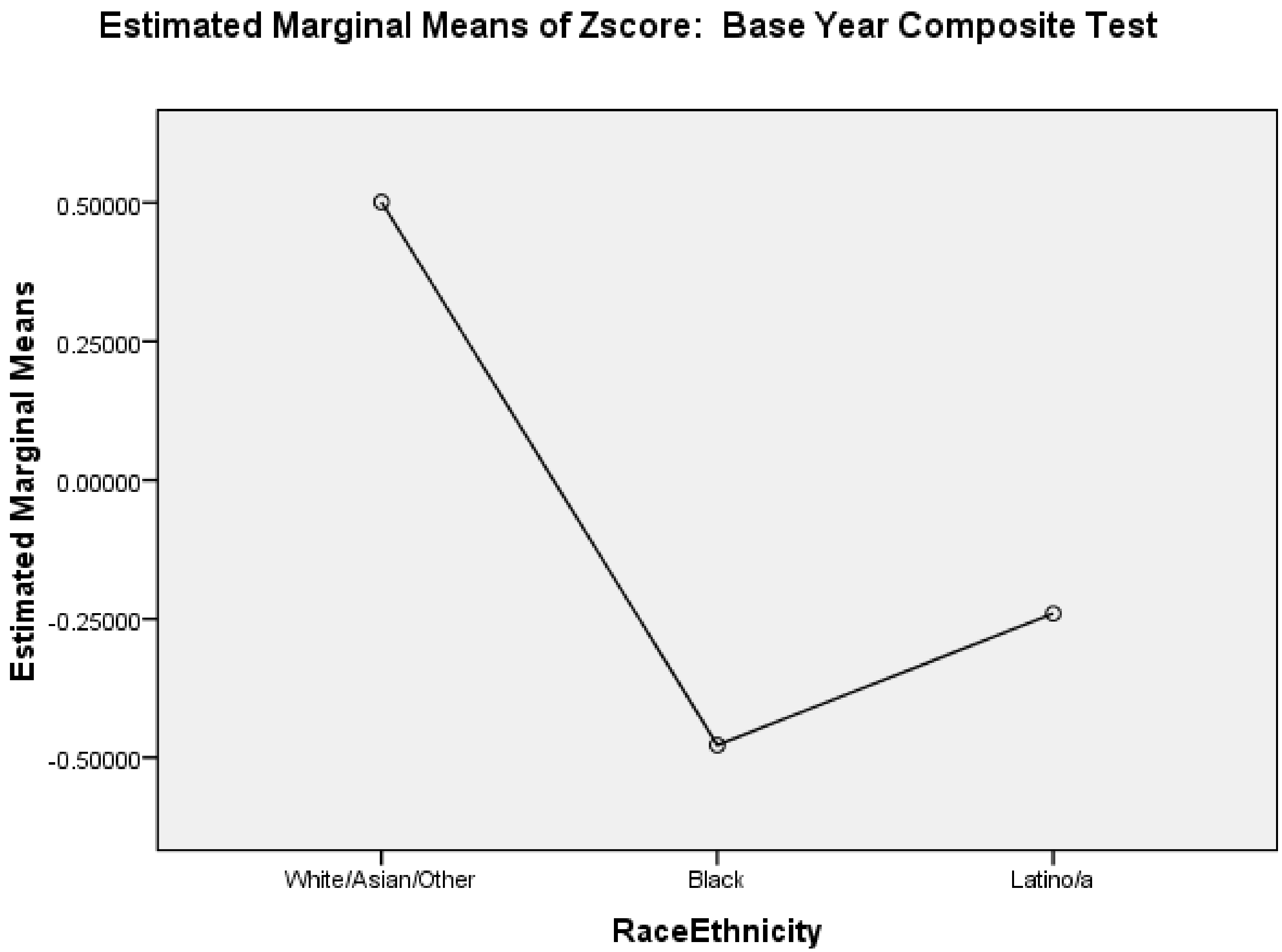
(I) RaceEthnicity	(J) RaceEthnicity	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
White/Asian/Other	Black	.9783362*	.06783237	.000	.8156840	1.1409885
	Latino/a	.7413527*	.06705305	.000	.5805692	.9021362
Black	White/Asian/Other	-.9783362*	.06783237	.000	-1.1409885	-.8156840
	Latino/a	-.2369835*	.07309953	.004	-.4122657	-.0617014
Latino/a	White/Asian/Other	-.7413527*	.06705305	.000	-.9021362	-.5805692
	Black	.2369835*	.07309953	.004	.0617014	.4122657

Based on observed means.
The error term is Mean Square(Error) = .815.

*. The mean difference is significant at the 0.05 level.

High School and Beyond (HSB.sav)





Understanding Causes of Illness (ILLCAUSE.sav)

- **Overview: Data for investigating differences in children's understanding of the causes of illness, by their health status.**
- **Source: Perrin E.C., Sayer A.G., and Willett J.B. (1991).**
  Sticks And Stones May Break My Bones: Reasoning About Illness Causality And Body Functioning In Children Who Have A Chronic Illness, *Pediatrics*, 88(3), 608-19.
- **Sample: 301 children, including a sub-sample of 205 who were described as asthmatic, diabetic,or healthy. After further reductions due to the *list-wise deletion* of cases with missing data on one or more variables, the analytic sub-sample used in class ends up containing: 33 diabetic children, 68 asthmatic children and 93 healthy children.**
- **Variables:**

(IllCause)	A Measure of Understanding of Illness Causality
(SocioEconomicStatus)	1=Low SES, 2=Lower Middle, 3=Upper Middle 4 = High SES
(HealthStatus)	1=Healthy, 2=Asthmatic 3=Diabetic

Dummy Variables for SocioEconomicStatus:

(LowSES)	1=Low SES, 0=Else
(LowerMiddleSES)	1=Lower MiddleSES, 0=Else
(HighSES)	1=High SES, 0=Else

*Note that we will use SocioEconomicStatus=3, Upper Middle SES, as our reference category.

Dummy Variables for HealthStatus:

(Asthmatic)	1=Asthmatic, 0=Else
(Diabetic)	1=Diabetic, 0=Else

*Note that we will use HealthStatus=1, Healthy, as our reference category.

Understanding Causes of Illness (ILLCAUSE.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.268[a]	.072	.057	.99236

a. Predictors: (Constant), HighSES, LowSES, LowerMiddleSES

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14.439	3	4.813	4.887	.003[a]
	Residual	187.108	190	.985		
	Total	201.547	193			

a. Predictors: (Constant), HighSES, LowSES, LowerMiddleSES

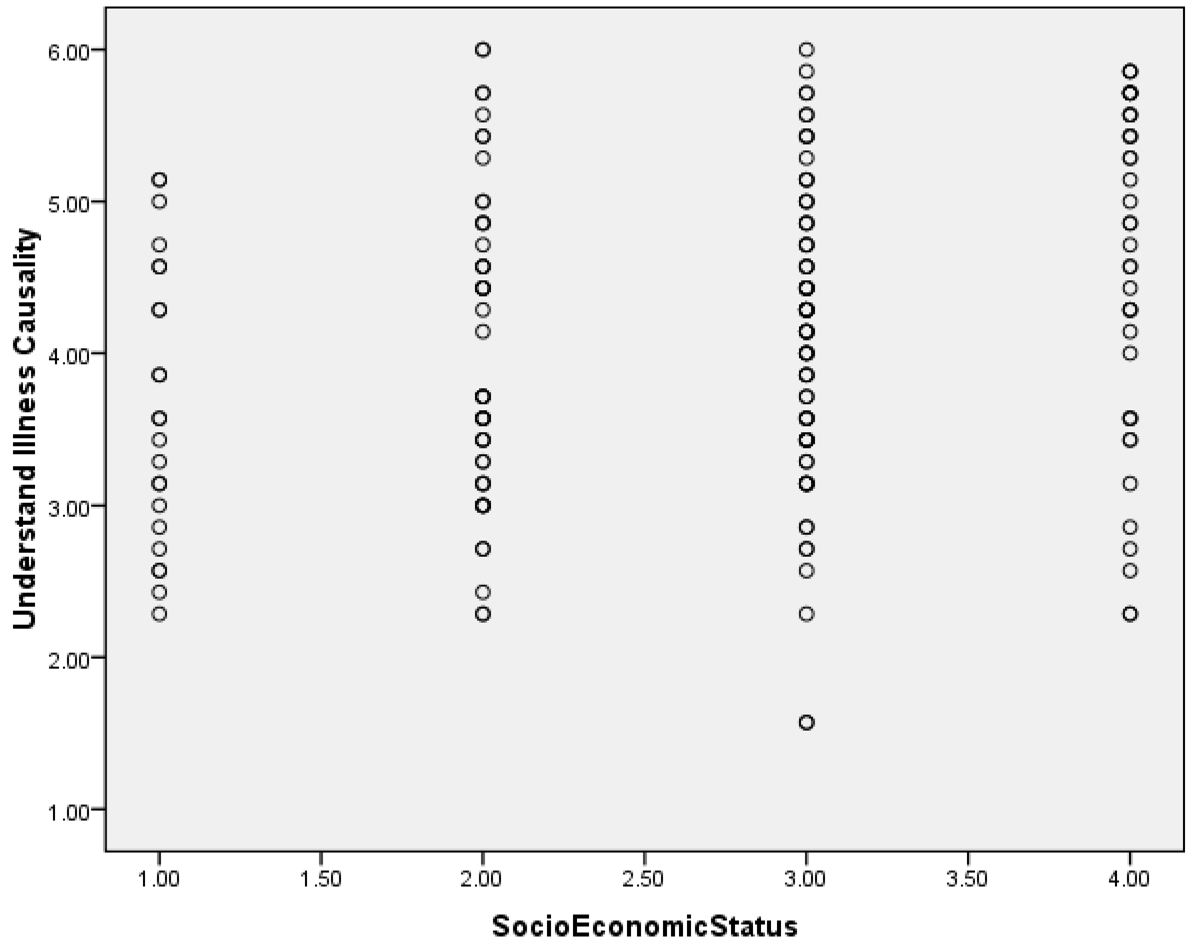
b. Dependent Variable: Understand Illness Causality

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.114	.111		37.083	.000	3.895	4.333
	LowSES	-.462	.235	-.147	-1.969	.050	-.925	.001
	LowerMiddleSES	-.100	.179	-.043	-.559	.577	-.453	.253
	HighSES	.471	.191	.189	2.471	.014	.095	.847

a. Dependent Variable: Understand Illness Causality

Understanding Causes of Illness (ILLCAUSE.sav)



Understanding Causes of Illness (ILLCAUSE.sav)

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
SocioEconomicStatus	1	Low SES	23
	2	Lower Middle SES	50
	3	Upper Midde SES	80
	4	High SES	41

Tests of Between-Subjects Effects

Dependent Variable:Understand Illness Causality

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	14.439[a]	3	4.813	4.887	.003
Intercept	2668.643	1	2668.643	2709.892	.000
SocioEconomicStatus	14.439	3	4.813	4.887	.003
Error	187.108	190	.985		
Total	3515.849	194			
Corrected Total	201.547	193			

a. R Squared = .072 (Adjusted R Squared = .057)

Understanding Causes of Illness (ILLCAUSE.sav)

Contrast Results (K Matrix)

SocioEconomicStatus Simple Contrast[a]			Dependent ... Understand Illness Causality
Level 1 vs. Level 3	Contrast Estimate		-.462
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.462
	Std. Error		.235
	Sig.		.050
	95% Confidence Interval for Difference	Lower Bound	-.925
		Upper Bound	.001
Level 2 vs. Level 3	Contrast Estimate		-.100
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.100
	Std. Error		.179
	Sig.		.577
	95% Confidence Interval for Difference	Lower Bound	-.453
		Upper Bound	.253
Level 4 vs. Level 3	Contrast Estimate		.471
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		.471
	Std. Error		.191
	Sig.		.014
	95% Confidence Interval for Difference	Lower Bound	.095
		Upper Bound	.847

a. Reference category = 3

Understanding Causes of Illness (ILLCAUSE.sav)

Multiple Comparisons

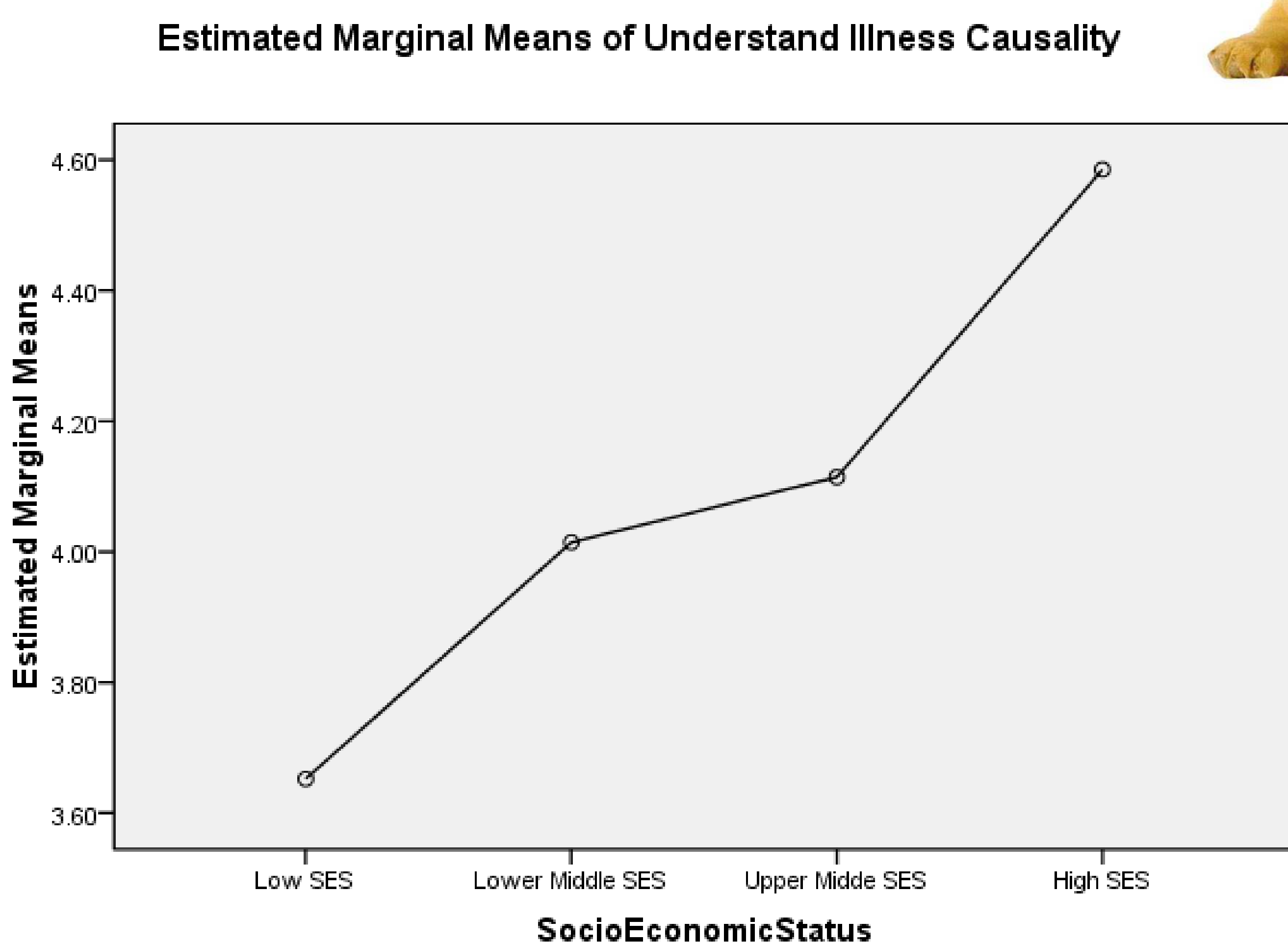
Understand Illness Causality
Bonferroni

(I) SocioEconomic Status	(J) SocioEconomic Status	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Low SES	Lower Middle SES	-.3621	.25002	.895	-1.0288	.3045
	Upper Midde SES	-.4622	.23479	.303	-1.0882	.1638
	High SES	-.9332*	.25853	.002	-1.6225	-.2439
Lower Middle SES	Low SES	.3621	.25002	.895	-.3045	1.0288
	Upper Midde SES	-.1000	.17890	1.000	-.5770	.3769
	High SES	-.5710*	.20908	.041	-1.1285	-.0136
Upper Midde SES	Low SES	.4622	.23479	.303	-.1638	1.0882
	Lower Middle SES	.1000	.17890	1.000	-.3769	.5770
	High SES	-.4710	.19060	.086	-.9792	.0372
High SES	Low SES	.9332*	.25853	.002	.2439	1.6225
	Lower Middle SES	.5710*	.20908	.041	.0136	1.1285
	Upper Midde SES	.4710	.19060	.086	-.0372	.9792

Based on observed means.
The error term is Mean Square(Error) = .985.

*. The mean difference is significant at the 0.05 level.

Understanding Causes of Illness (ILLCAUSE.sav)



Understanding Causes of Illness (ILLCAUSE.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.444[a]	.197	.189	.92042

a. Predictors: (Constant), Diabetic, Asthmatic

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	39.737	2	19.869	23.453	.000[a]
	Residual	161.810	191	.847		
	Total	201.547	193			

a. Predictors: (Constant), Diabetic, Asthmatic

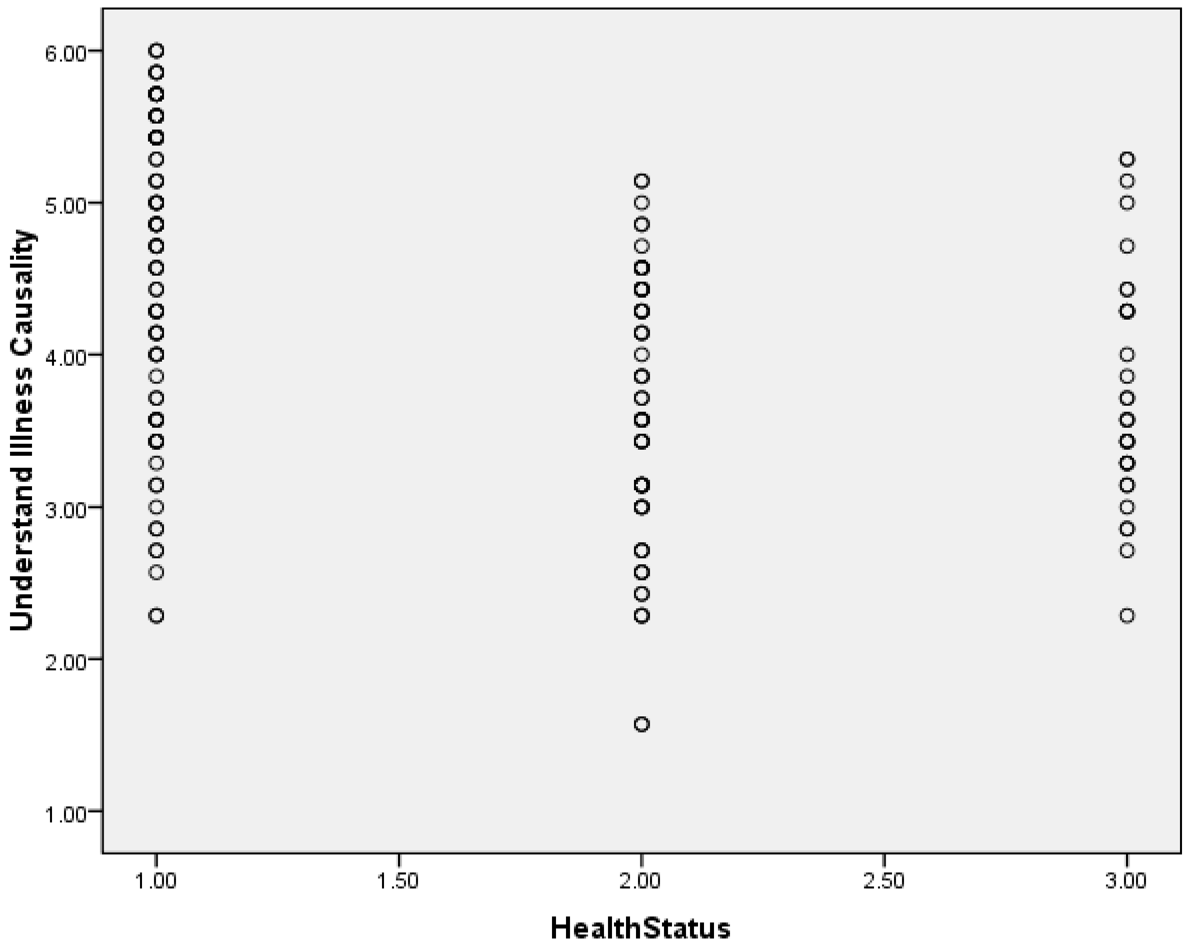
b. Dependent Variable: Understand Illness Causality

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.604	.095		48.235	.000	4.415	4.792
	Asthmatic	-.936	.147	-.438	-6.371	.000	-1.225	-.646
	Diabetic	-.837	.186	-.309	-4.490	.000	-1.205	-.469

a. Dependent Variable: Understand Illness Causality

Understanding Causes of Illness (ILLCAUSE.sav)



Understanding Causes of Illness (ILLCAUSE.sav)

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
HealthStatus	1	Healthy	93
	2	Asthmatic	68
	3	Diabetic	33

Tests of Between-Subjects Effects

Dependent Variable:Understand Illness Causality

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	39.737[a]	2	19.869	23.453	.000
Intercept	2598.824	1	2598.824	3067.647	.000
HealthStatus	39.737	2	19.869	23.453	.000
Error	161.810	191	.847		
Total	3515.849	194			
Corrected Total	201.547	193			

a. R Squared = .197 (Adjusted R Squared = .189)

Understanding Causes of Illness (ILLCAUSE.sav)

Contrast Results (K Matrix)

HealthStatus Helmert Contrast			Dependent ... Understand Illness Causality
Level 1 vs. Later	Contrast Estimate		.886
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		.886
	Std. Error		.137
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	.617
		Upper Bound	1.156
Level 2 vs. Level 3	Contrast Estimate		-.098
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.098
	Std. Error		.195
	Sig.		.615
	95% Confidence Interval for Difference	Lower Bound	-.483
		Upper Bound	.287

Understanding Causes of Illness (ILLCAUSE.sav)

Multiple Comparisons

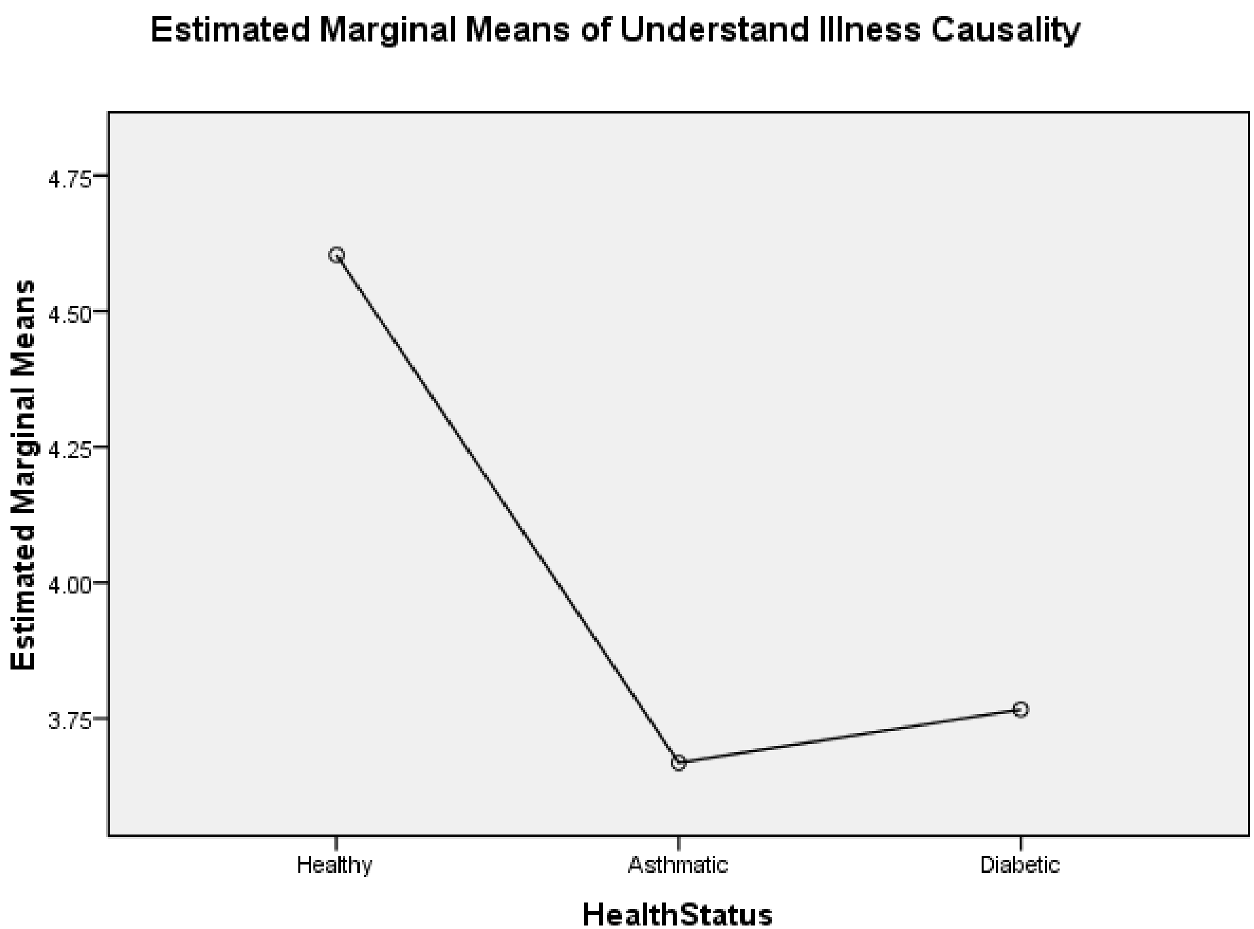
Understand Illness Causality
Bonferroni

(I) Health Status	(J) Health Status	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Healthy	Asthmatic	.9356*	.14686	.000	.5809	1.2903
	Diabetic	.8373*	.18650	.000	.3869	1.2878
Asthmatic	Healthy	-.9356*	.14686	.000	-1.2903	-.5809
	Diabetic	-.0983	.19527	1.000	-.5699	.3734
Diabetic	Healthy	-.8373*	.18650	.000	-1.2878	-.3869
	Asthmatic	.0983	.19527	1.000	-.3734	.5699

Based on observed means.
The error term is Mean Square(Error) = .847.

*. The mean difference is significant at the 0.05 level.

Understanding Causes of Illness (ILLCAUSE.sav)



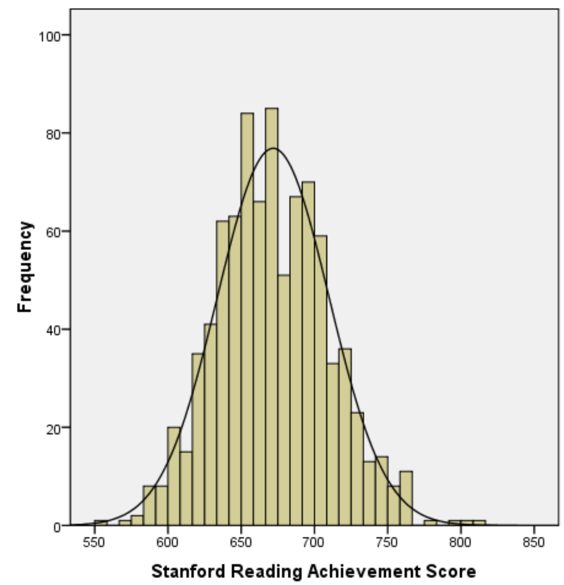
Children of Immigrants (ChildrenOfImmigrants.sav)

- **Overview: "CILS is a longitudinal study designed to study the adaptation process of the immigrant second generation which is defined broadly as U.S.-born children with at least one foreign-born parent or children born abroad but brought at an early age to the United States. The original survey was conducted with large samples of second-generation children attending the 8th and 9th grades in public and private schools in the metropolitan areas of Miami/Ft. Lauderdale in Florida and San Diego, California" (from the website description of the data set).**
- **Source: Portes, Alejandro, & Ruben G. Rumbaut (2001). *Legacies: The Story of the Immigrant SecondGeneration*. Berkeley CA: University of California Press.**
- **Sample: Random sample of 880 participants obtained through the website.**
- **Variables:**

(Reading)	Stanford Reading Achievement Scores
(Depressed)	1=The Student is Depressed, 0=Not Depressed
(SESCat)	A Relative Measure Of Socio-Economic Status 1=Low SES, 2=Mid SES, 3=High SES

Dummy Variables for SESCat:

(LowSES)	1=Low SES, 0=Else
(MidSES)	1=Mid SES, 0=Else
(HighSES)	1=High SES, 0=Else



Children of Immigrants (ChildrenOfImmigrants.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.108[a]	.012	.011	37.851

a. Predictors: (Constant), Depressed

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14974.978	1	14974.978	10.452	.001[a]
	Residual	1257919.200	878	1432.710		
	Total	1272894.177	879			

a. Predictors: (Constant), Depressed

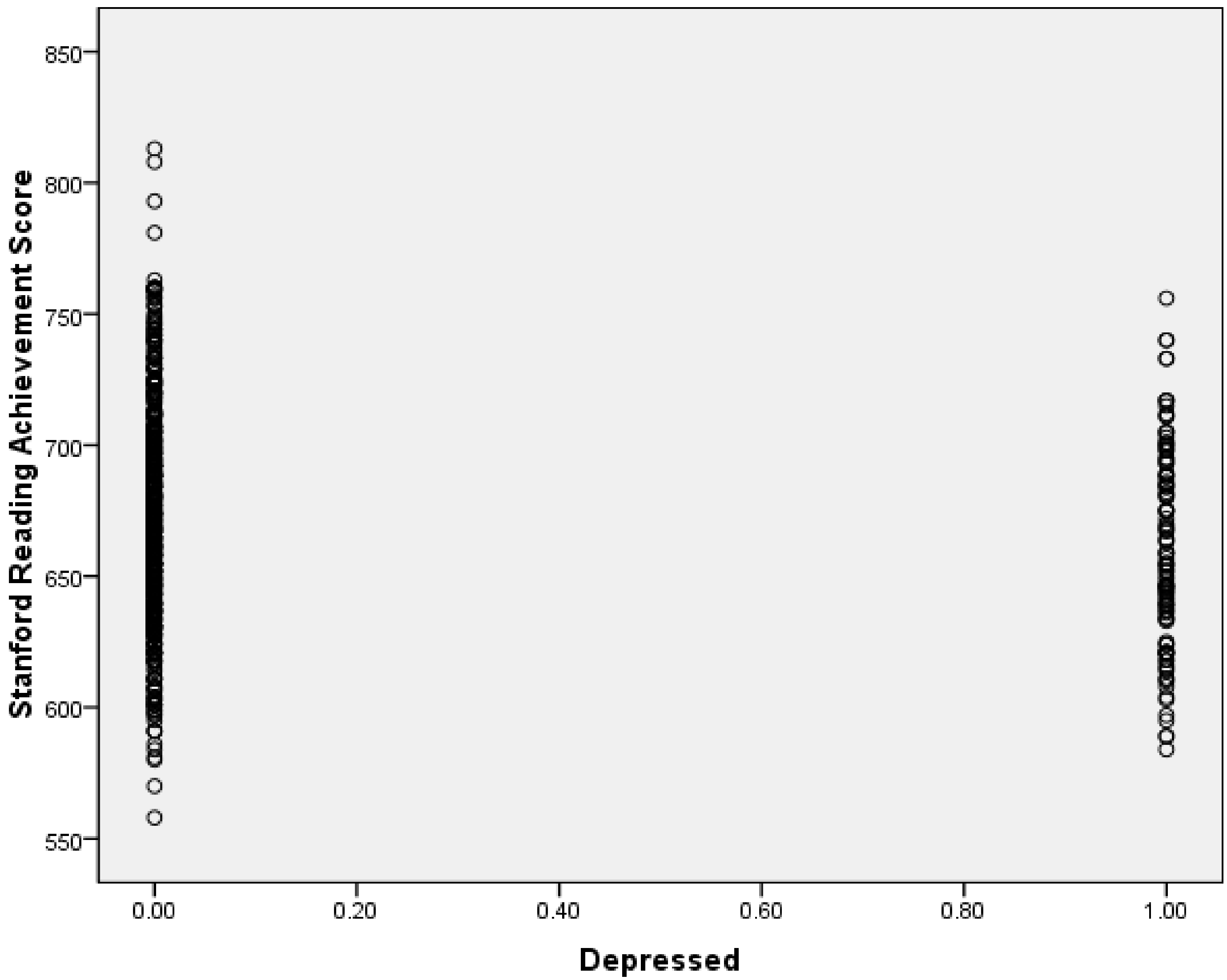
b. Dependent Variable: Stanford Reading Achievement Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	673.407	1.368		492.394	.000	670.723	676.091
	Depressed	-12.285	3.800	-.108	-3.233	.001	-19.742	-4.827

a. Dependent Variable: Stanford Reading Achievement Score

Children of Immigrants (ChildrenOfImmigrants.sav)



Univariate Analysis of Variance

Between-Subjects Factors

		N
Depressed	0	766
	1	114

Tests of Between-Subjects Effects

Dependent Variable:Stanford Reading Achievement Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	14974.978[a]	1	14974.978	10.452	.001
Intercept	1.767E8	1	1.767E8	123352.932	.000
Depressed	14974.978	1	14974.978	10.452	.001
Error	1257919.200	878	1432.710		
Total	3.984E8	880			
Corrected Total	1272894.177	879			

a. R Squared = .012 (Adjusted R Squared = .011)

Children of Immigrants (ChildrenOfImmigrants.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.332[a]	.110	.108	35.941

a. Predictors: (Constant), HighSES, LowSES

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	140028.573	2	70014.287	54.201	.000[a]
	Residual	1132865.604	877	1291.751		
	Total	1272894.177	879			

a. Predictors: (Constant), HighSES, LowSES

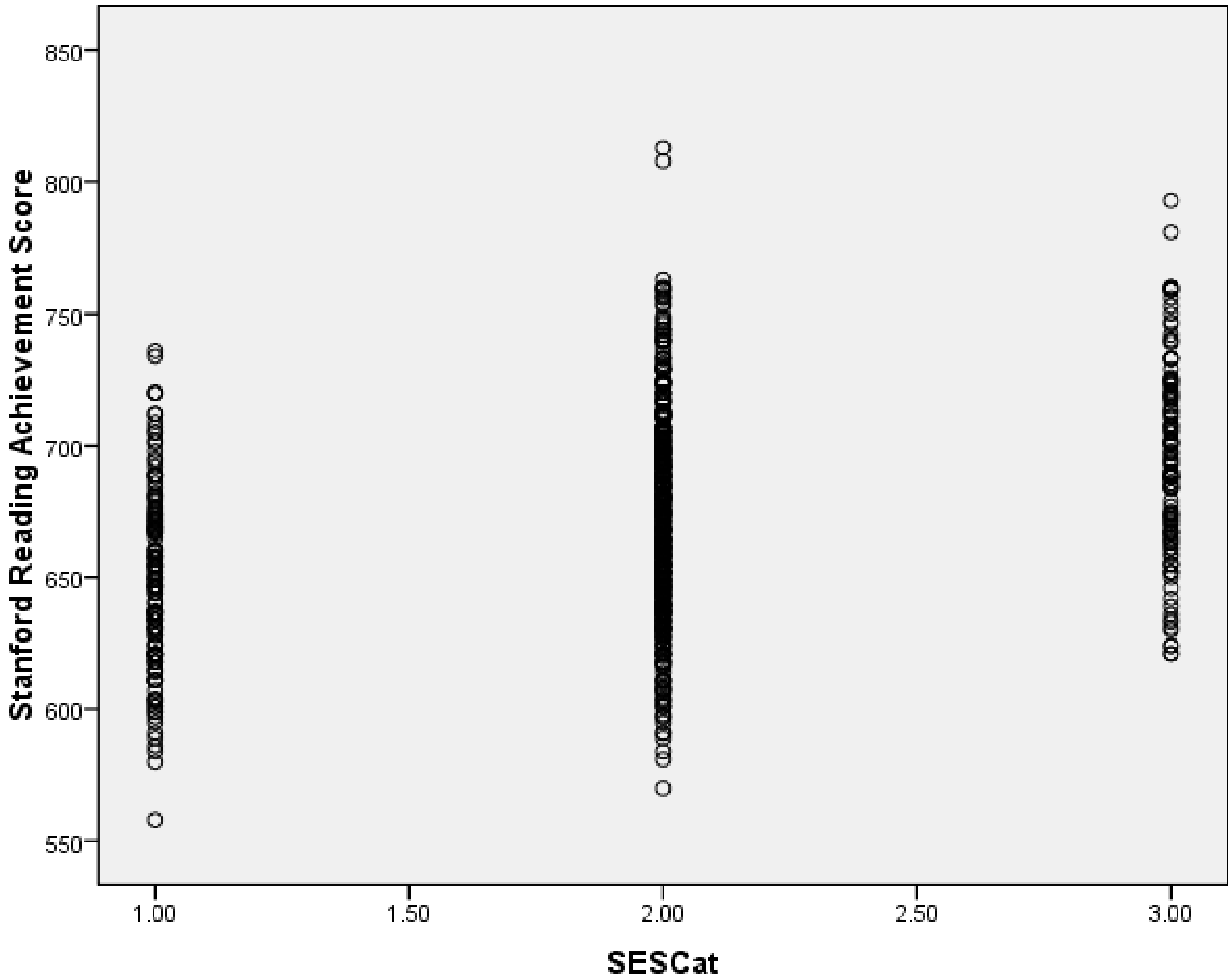
b. Dependent Variable: Stanford Reading Achievement Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	672.062	1.454		462.211	.000	669.208	674.916
	LowSES	-22.861	3.377	-.219	-6.769	.000	-29.490	-16.232
	HighSES	22.776	3.471	.212	6.561	.000	15.963	29.590

a. Dependent Variable: Stanford Reading Achievement Score

Children of Immigrants (ChildrenOfImmigrants.sav)



Children of Immigrants (ChildrenOfImmigrants.sav)

Univariate Analysis of Variance

Between-Subjects Factors

		N
SESCat	1	139
	2	611
	3	130

Tests of Between-Subjects Effects

Dependent Variable:Stanford Reading Achievement Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	140028.573[a]	2	70014.287	54.201	.000
Intercept	2.460E8	1	2.460E8	190437.174	.000
SESCat	140028.573	2	70014.287	54.201	.000
Error	1132865.604	877	1291.751		
Total	3.984E8	880			
Corrected Total	1272894.177	879			

a. R Squared = .110 (Adjusted R Squared = .108)

Children of Immigrants (ChildrenOfImmigrants.sav)

Contrast Results (K Matrix)

SESCat Simple Contrast[a]			Dependent ... Stanford Reading Achievement Score
Level 1 vs. Level 2	Contrast Estimate		-22.861
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-22.861
	Std. Error		3.377
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-29.490
		Upper Bound	-16.232
Level 3 vs. Level 2	Contrast Estimate		22.776
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		22.776
	Std. Error		3.471
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	15.963
		Upper Bound	29.590

a. Reference category = 2

Children of Immigrants (ChildrenOfImmigrants.sav)

Multiple Comparisons

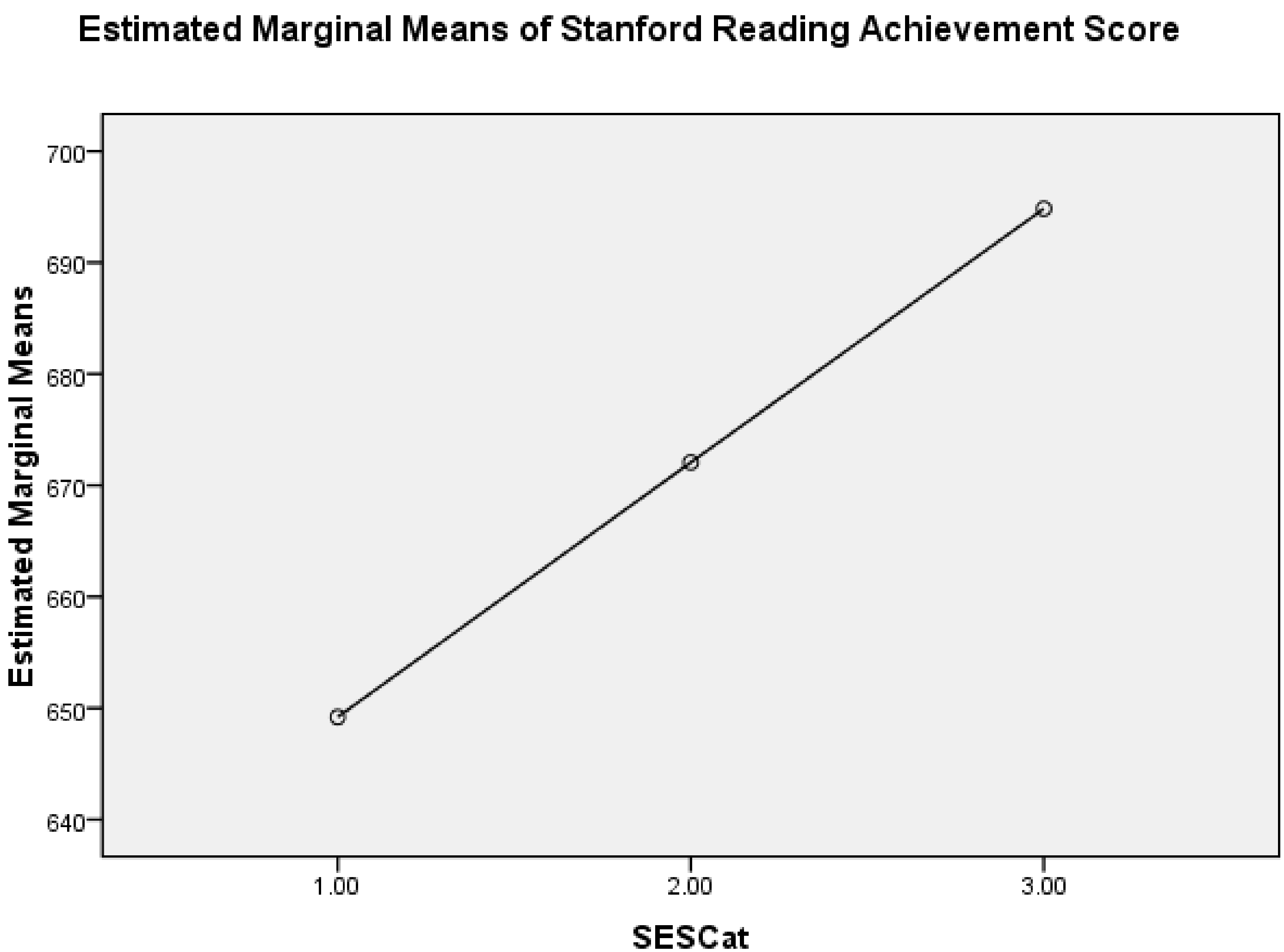
Stanford Reading Achievement Score
Bonferroni

(I) SES Cat	(J) SES Cat	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	2	-22.86*	3.377	.000	-30.96	-14.76
	3	-45.64*	4.385	.000	-56.16	-35.12
2	1	22.86*	3.377	.000	14.76	30.96
	3	-22.78*	3.471	.000	-31.10	-14.45
3	1	45.64*	4.385	.000	35.12	56.16
	2	22.78*	3.471	.000	14.45	31.10

Based on observed means.
The error term is Mean Square(Error) = 1291.751.

*. The mean difference is significant at the 0.05 level.

Children of Immigrants (ChildrenOfImmigrants.sav)



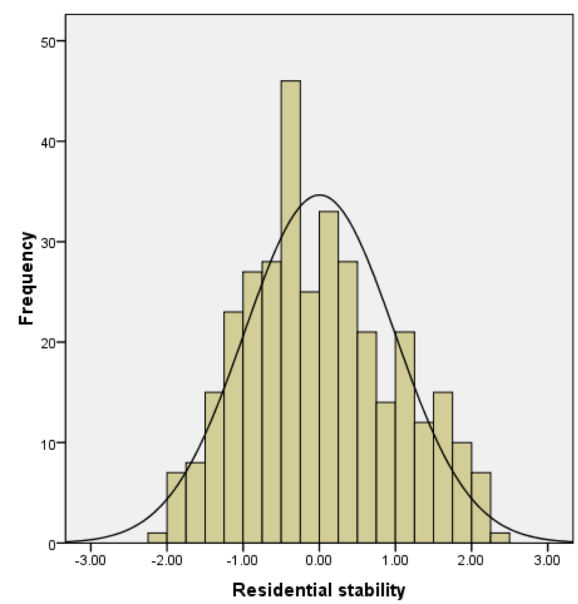
Human Development in Chicago Neighborhoods (Neighborhoods.sav)

- **These data were collected as part of the Project on Human Development in Chicago Neighborhoods in 1995.**
- **Source: Sampson, R.J., Raudenbush, S.W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277, 918-924.**
- **Sample: The data described here consist of information from 343 Neighborhood Clusters in Chicago Illinois. Some of the variables were obtained by project staff from the 1990 Census and city records. Other variables were obtained through questionnaire interviews with 8782 Chicago residents who were interviewed in their homes.**
- **Variables:**

(ResStab)	Residential Stability, A Measure Of Neighborhood Flux
(NoMurder95)	1=No Murders in 1995, 0=At Least One Murder in 1995
(SES)	A Relative Measure Of Socio-Economic Status 1=Low SES, 2=Mid SES, 3=High SES

Dummy Variables for MothEdCat:

(LowSES)	1=Low SES, 0=Else
(MidSES)	1=Mid SES, 0=Else
(HighSES)	1=High SES, 0=Else



Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.035[a]	.001	-.002	.98511

a. Predictors: (Constant), NoMurder95

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.413	1	.413	.426	.514[a]
	Residual	329.947	340	.970		
	Total	330.361	341			

a. Predictors: (Constant), NoMurder95

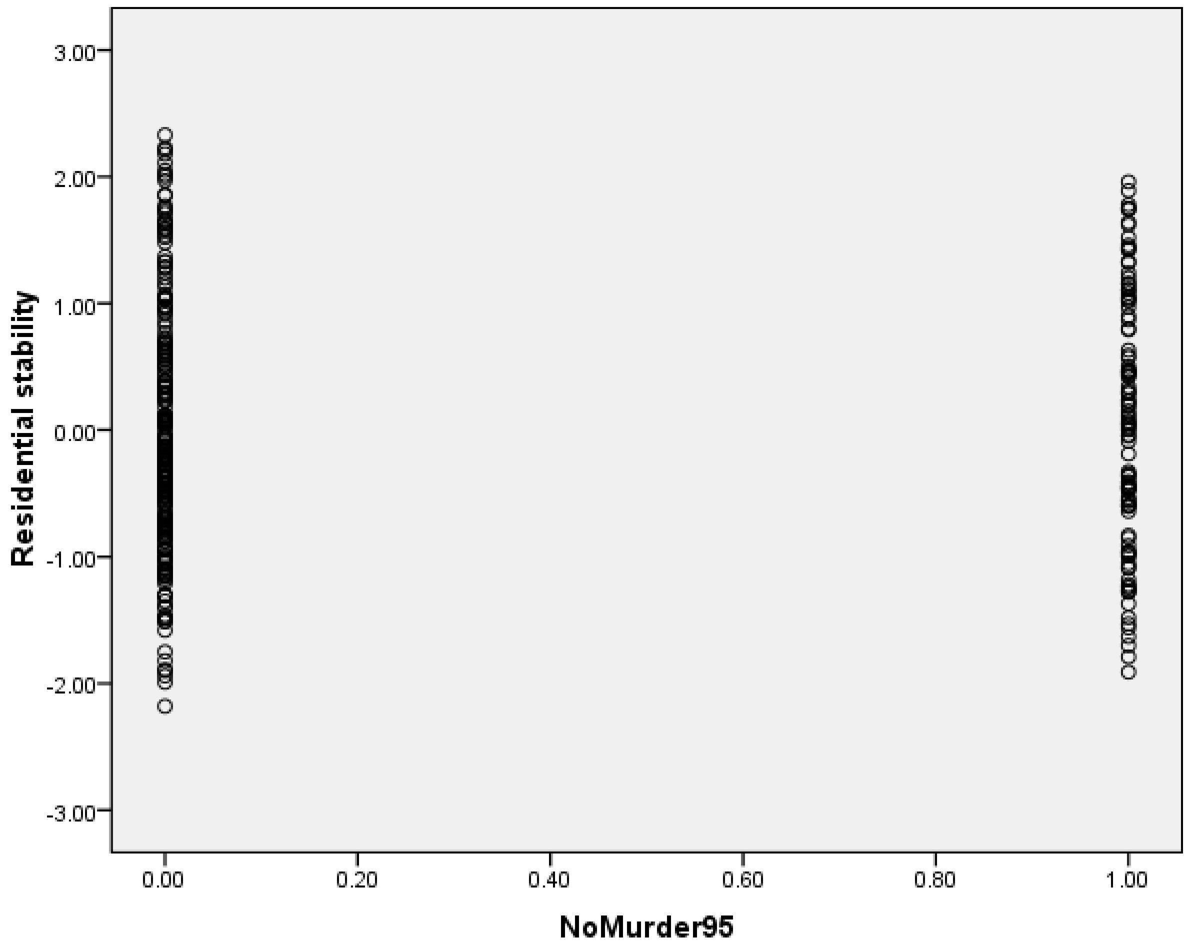
b. Dependent Variable: Residential stability

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.021	.065		-.329	.743	-.148	.106
	NoMurder95	.074	.114	.035	.653	.514	-.150	.299

a. Dependent Variable: Residential stability

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
NoMurder95	0	At Least One Murder in 1995	232
	1	No Murders in 1995	110

Tests of Between-Subjects Effects

Dependent Variable:Residential stability

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.413[a]	1	.413	.426	.514
Intercept	.076	1	.076	.078	.780
NoMurder95	.413	1	.413	.426	.514
Error	329.947	340	.970		
Total	330.363	342			
Corrected Total	330.361	341			

a. R Squared = .001 (Adjusted R Squared = -.002)

Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.029[a]	.001	-.005	.98675

a. Predictors: (Constant), HighSES, LowSES

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.282	2	.141	.145	.865[a]
	Residual	330.079	339	.974		
	Total	330.361	341			

a. Predictors: (Constant), HighSES, LowSES

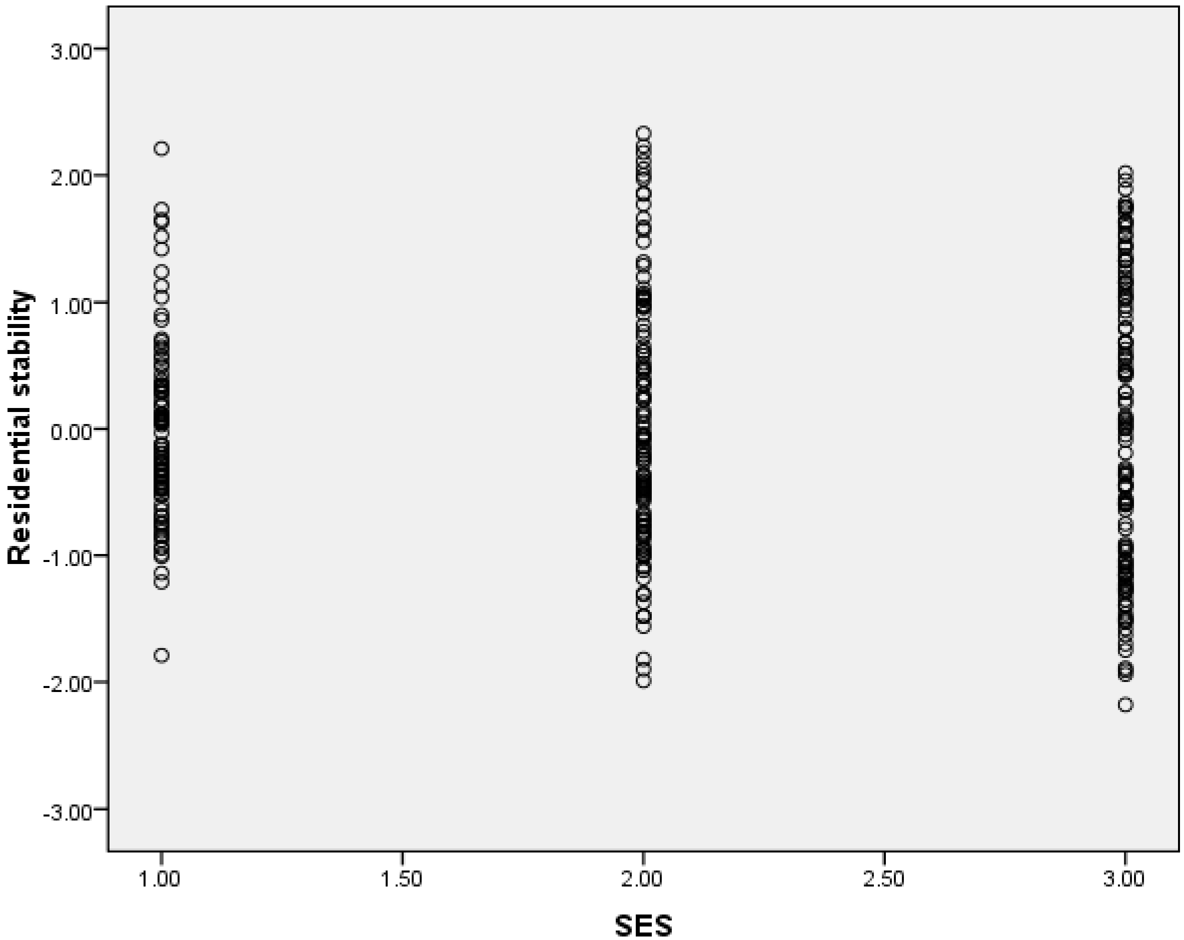
b. Dependent Variable: Residential stability

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.040	.090		.447	.655	-.136	.217
	LowSES	-.069	.134	-.032	-.511	.610	-.332	.195
	HighSES	-.049	.126	-.024	-.391	.696	-.298	.199

a. Dependent Variable: Residential stability

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
SES	1	Low SES	98
	2	Mid SES	121
	3	High SES	123

Tests of Between-Subjects Effects

Dependent Variable:Residential stability

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.282[a]	2	.141	.145	.865
Intercept	.000	1	.000	.000	.988
SES	.282	2	.141	.145	.865
Error	330.079	339	.974		
Total	330.363	342			
Corrected Total	330.361	341			

a. R Squared = .001 (Adjusted R Squared = -.005)

Human Development in Chicago Neighborhoods (Neighborhoods.sav)

Contrast Results (K Matrix)

SES Simple Contrast[a]			Dependent ... Residential stability
Level 1 vs. Level 2	Contrast Estimate		-.069
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.069
	Std. Error		.134
	Sig.		.610
	95% Confidence Interval for Difference	Lower Bound	-.332
		Upper Bound	.195
Level 3 vs. Level 2	Contrast Estimate		-.049
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.049
	Std. Error		.126
	Sig.		.696
	95% Confidence Interval for Difference	Lower Bound	-.298
		Upper Bound	.199

a. Reference category = 2

Human Development in Chicago Neighborhoods (Neighborhoods.sav)

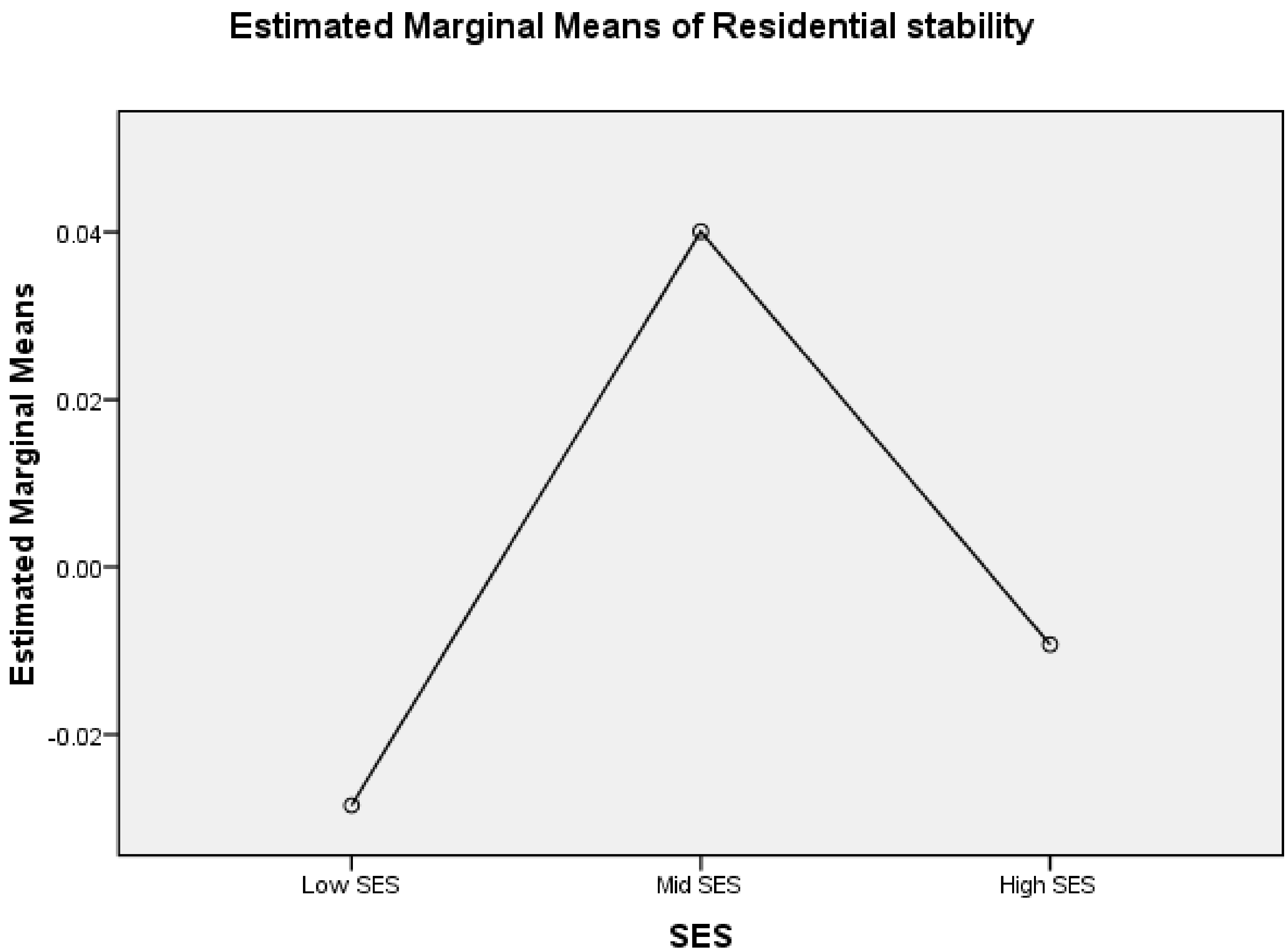
Multiple Comparisons

Residential stability
Bonferroni

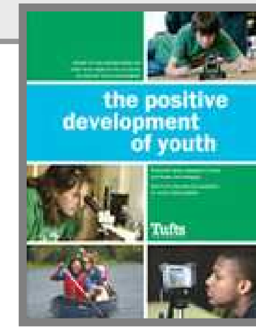
(I) SES	(J) SES	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Low SES	Mid SES	-.0686	.13410	1.000	-.3912	.2541
	High SES	-.0192	.13361	1.000	-.3407	.3023
Mid SES	Low SES	.0686	.13410	1.000	-.2541	.3912
	High SES	.0494	.12635	1.000	-.2546	.3533
High SES	Low SES	.0192	.13361	1.000	-.3023	.3407
	Mid SES	-.0494	.12635	1.000	-.3533	.2546

Based on observed means.
The error term is Mean Square(Error) = .974.

Human Development in Chicago Neighborhoods (Neighborhoods.sav)



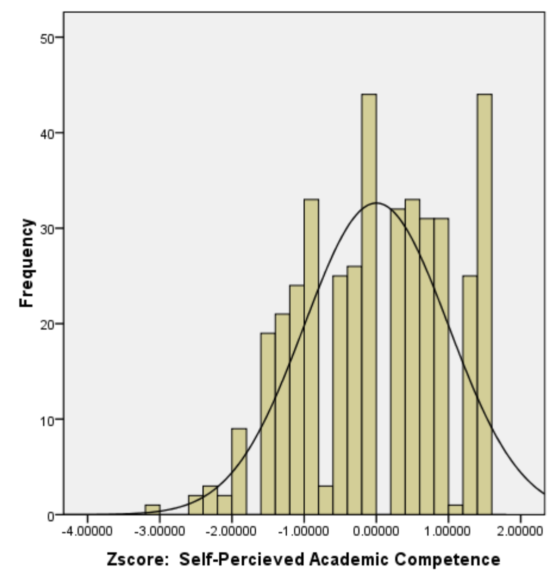
4-H Study of Positive Youth Development (4H.sav)



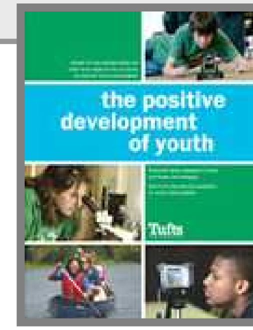
- **4-H Study of Positive Youth Development**
- **Source: Subset of data from IARYD, Tufts University**
- **Sample: These data consist of seventh graders who participated in Wave 3 of the 4-H Study of Positive Youth Development at Tufts University. This subfile is a substantially sampled-down version of the original file, as all the cases with any missing data on these selected variables were eliminated.**
- **Variables:**

(ZAcadComp) Standardized Self-Perceived Academic Competence
(SexFem) 1=Female, 0=Male
(MothEdCat) Mother's Educational Attainment Category
1=High School Dropout, 2=High School Graduate,
3 =Up To 3 Years of College, 4 = 4-Plus Years of College

Dummy Variables for MothEdCat:
(MomHSDropout) 1=High School Dropout, 0=Else
(MomHSGrad) 1=High School Graduate, 0=Else
(MomUpTo3YRSCollege) 1=Up To 3 Years of College, 0=Else
(Mom4plusYRSCollege) 1=4-Plus Years of College, 0=Else



4-H Study of Positive Youth Development (4H.sav)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.041[a]	.002	.000	1.00039421

a. Predictors: (Constant), Female = 1, Male = 0

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.679	1	.679	.679	.411[a]
	Residual	407.321	407	1.001		
	Total	408.000	408			

a. Predictors: (Constant), Female = 1, Male = 0

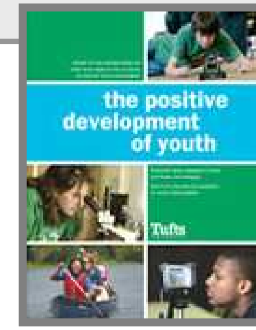
b. Dependent Variable: Zscore: Self-Percieved Academic Competence

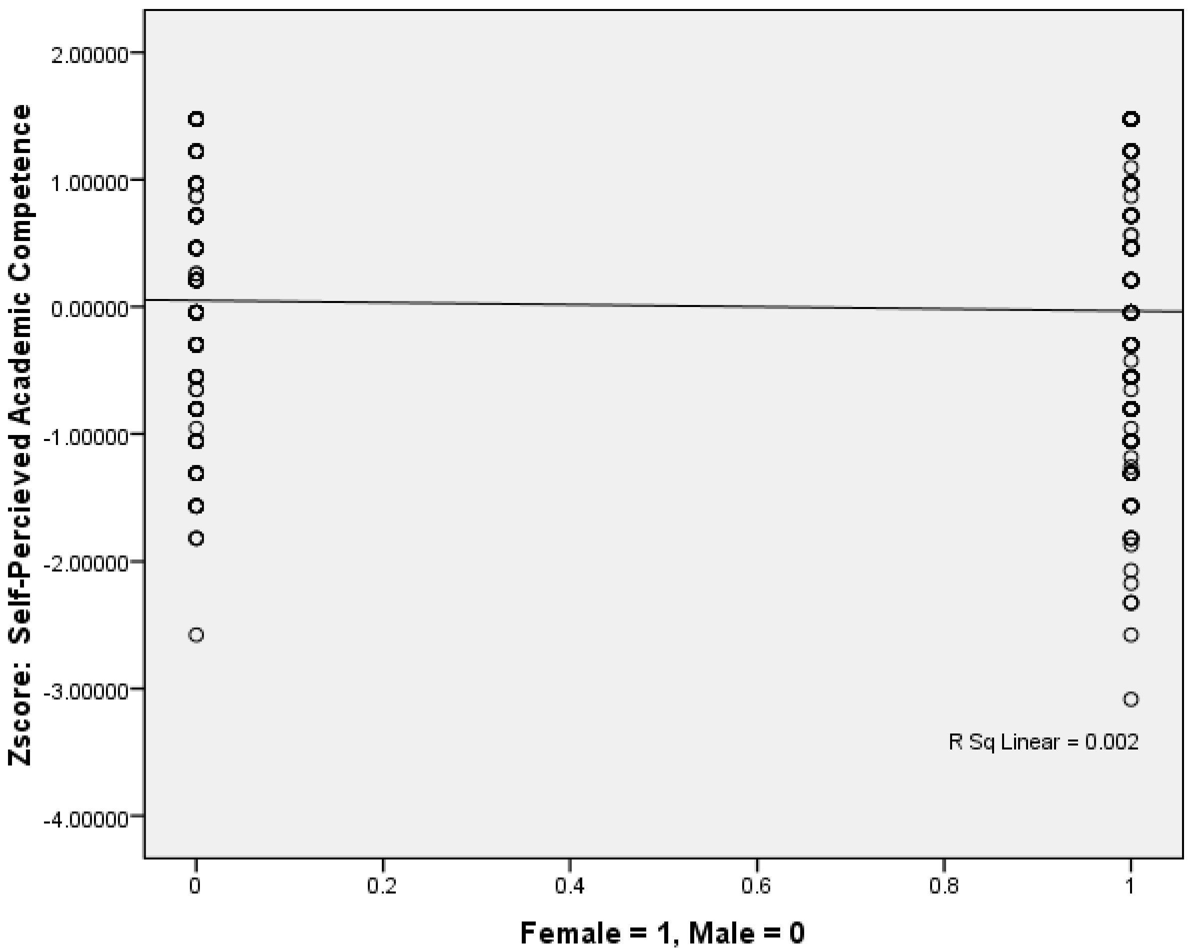
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.050	.078		.636	.525	-.104	.203
	Female = 1, Male = 0	-.083	.101	-.041	-.824	.411	-.281	.115

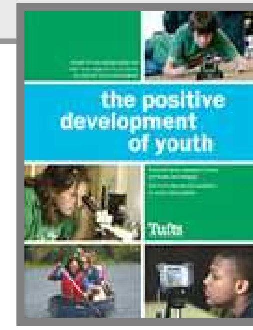
a. Dependent Variable: Zscore: Self-Percieved Academic Competence

4-H Study of Positive Youth Development (4H.sav)





4-H Study of Positive Youth Development (4H.sav)



Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
Female = 1, Male = 0	0	Boy	165
	1	Girl	244

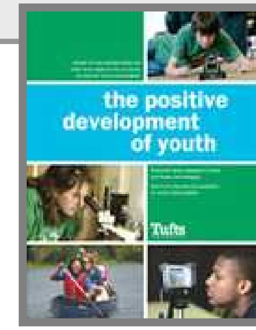
Tests of Between-Subjects Effects

Dependent Variable:Zscore: Self-Percieved Academic Competence

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.679[a]	1	.679	.679	.411
Intercept	.025	1	.025	.025	.874
SexFem	.679	1	.679	.679	.411
Error	407.321	407	1.001		
Total	408.000	409			
Corrected Total	408.000	408			

a. R Squared = .002 (Adjusted R Squared = -.001)

4-H Study of Positive Youth Development (4H.sav)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.333[a]	.111	.104	.94643681

a. Predictors: (Constant), Mom4plusYRSCollege, MomHSDropout, MomUpTo3YRSCollege

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45.224	3	15.075	16.829	.000[a]
	Residual	362.776	405	.896		
	Total	408.000	408			

a. Predictors: (Constant), Mom4plusYRSCollege, MomHSDropout, MomUpTo3YRSCollege

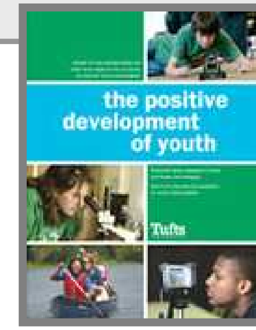
b. Dependent Variable: Zscore: Self-Percieved Academic Competence

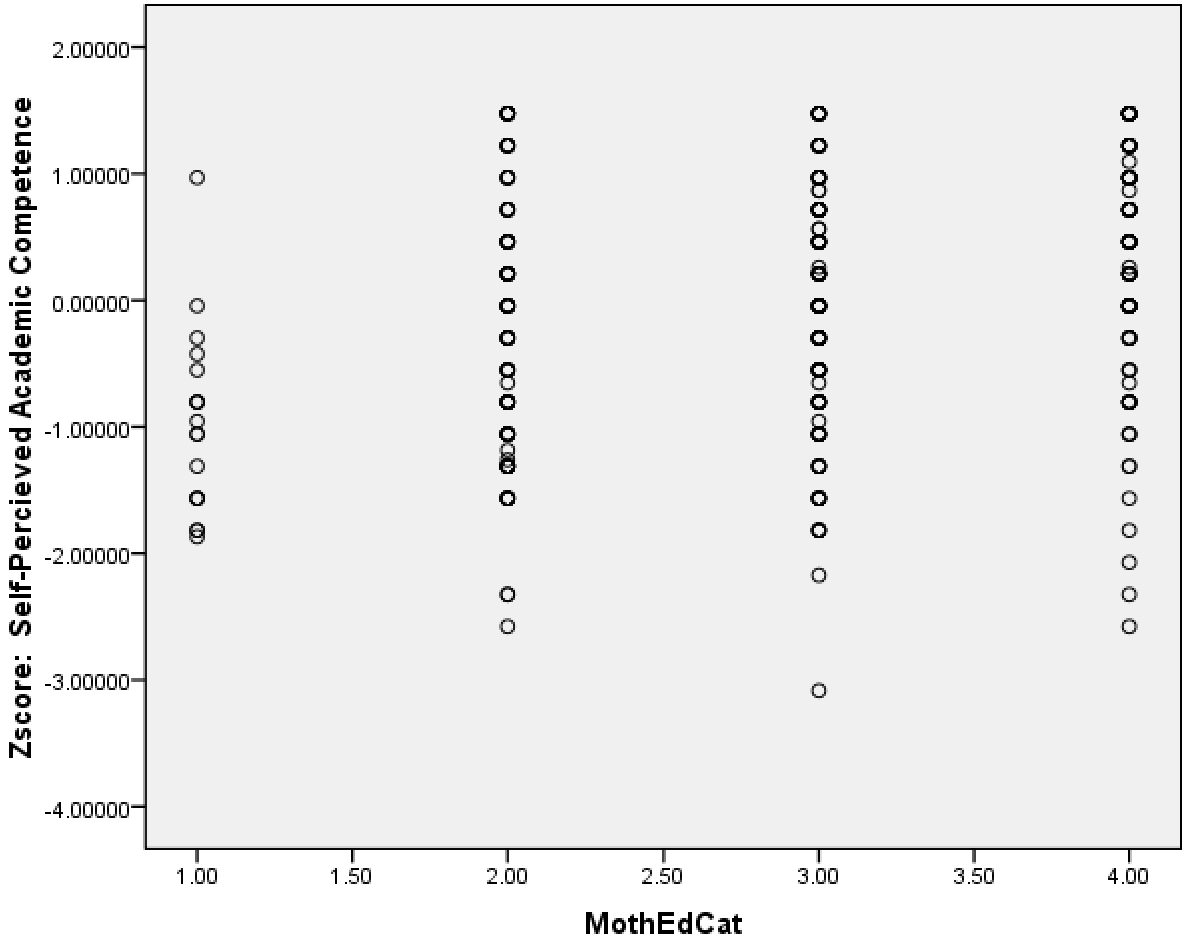
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.226	.090		-2.501	.013	-.403	-.048
	MomHSDropout	-.738	.241	-.152	-3.066	.002	-1.211	-.265
	MomUpTo3YRSCollege	.158	.118	.077	1.342	.180	-.073	.390
	Mom4plusYRSCollege	.647	.124	.298	5.230	.000	.404	.890

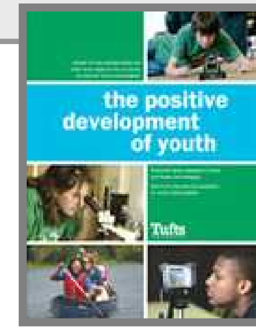
a. Dependent Variable: Zscore: Self-Percieved Academic Competence

4-H Study of Positive Youth Development (4H.sav)





4-H Study of Positive Youth Development (4H.sav)



Between-Subjects Factors

		Value Label	N
MothEdCat	1	Mom HS Dropout	18
	2	Mom HS Grad	110
	3	Mom Up to 3 YRS College	156
	4	Mom 4+ YRS College	125

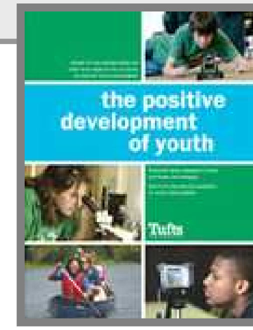
Tests of Between-Subjects Effects

Dependent Variable:Zscore: Self-Percieved Academic Competence

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	45.224[a]	3	15.075	16.829	.000
Intercept	8.819	1	8.819	9.846	.002
MothEdCat	45.224	3	15.075	16.829	.000
Error	362.776	405	.896		
Total	408.000	409			
Corrected Total	408.000	408			

a. R Squared = .111 (Adjusted R Squared = .104)

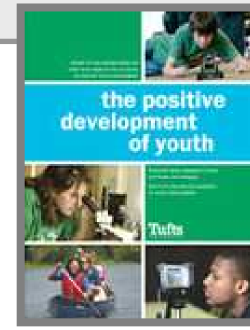
4-H Study of Positive Youth Development (4H.sav)



Contrast Results (K Matrix)

MothEdCat Helmert Contrast			Dependent ... Zscore: Self-Percieved Academic Competence
Level 1 vs. Later	Contrast Estimate		-1.006
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-1.006
	Std. Error		.228
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-1.455
		Upper Bound	-.557
Level 2 vs. Later	Contrast Estimate		-.403
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.403
	Std. Error		.107
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-.612
		Upper Bound	-.193
Level 3 vs. Level 4	Contrast Estimate		-.489
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.489
	Std. Error		.114
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-.712
		Upper Bound	-.266

4-H Study of Positive Youth Development (4H.sav)



Multiple Comparisons

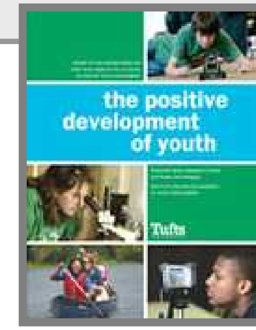
Zscore: Self-Percieved Academic Competence
Bonferroni

(I) MothEdCat	(J) MothEdCat	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Mom HS Dropout	Mom HS Grad	-.7377445*	.24063789	.014	-1.3757432	-.0997458
	Mom Up to 3 YRS College	-.8959237*	.23559588	.001	-1.5205546	-.2712927
	Mom 4+ YRS College	-1.3848849*	.23859887	.000	-2.0174776	-.7522922
Mom HS Grad	Mom HS Dropout	.7377445*	.24063789	.014	.0997458	1.3757432
	Mom Up to 3 YRS College	-.1581792	.11783486	1.000	-.4705926	.1542342
	Mom 4+ YRS College	-.6471405*	.12372977	.000	-.9751829	-.3190981
Mom Up to 3 YRS College	Mom HS Dropout	.8959237*	.23559588	.001	.2712927	1.5205546
	Mom HS Grad	.1581792	.11783486	1.000	-.1542342	.4705926
	Mom 4+ YRS College	-.4889613*	.11361286	.000	-.7901809	-.1877416
Mom 4+ YRS College	Mom HS Dropout	1.3848849*	.23859887	.000	.7522922	2.0174776
	Mom HS Grad	.6471405*	.12372977	.000	.3190981	.9751829
	Mom Up to 3 YRS College	.4889613*	.11361286	.000	.1877416	.7901809

Based on observed means.
The error term is Mean Square(Error) = .896.

*. The mean difference is significant at the 0.05 level.

4-H Study of Positive Youth Development (4H.sav)



Estimated Marginal Means of Zscore: Self-Percieved Academic Competence